LSE Career Accelerator in Data Analytics

Advanced Analytics for Organisational Impact

Course 3: Predicting Future Outcomes

Saima Abbas

30th September 2024

Contents

Business Problem: 4

Analytical Approach: 4

Data Wrangling: 4

Statistical Analysis: 4

Regression Analysis: 5

K-Means Clustering: 5

Sentiment Analysis: 5

Insights and Recommendation: 5

Customer Loyalty Insights 5

1. Descriptive Analysis 5

2. Grouped Loyalty Analysis: 6

3. Correlation Analysis: (Appendix 3) 7

4. Regression Models Overview 7

5. Decision Tree Regressor: (Appendix 9) 9

6. K-Means Clustering: (Appendix 10) 10

6. Sentiment Analysis of Customers reviews: (Appendix 11) 10

7. Recommendations: 11

Appendix 1: The 5 Whys Technique 13

Root Cause: 13

Solution Ideas Based on Root Cause: 13

Appendix 2: Regression Summaries Tables 14

Appendix 3: Correlation Matrix 19

19

Appendix 4: Histograms, Boxplots 20

Appendix 5: Average Loyalty Points earned by Customers based on Age, Education, Gender, Spending, Product and Salary. 26

Appendix 6: Scatterplots showing Correlation between different Predictor variables (Age, Salary, Spending, Product) and Target variable (Loyalty). 31

Appendix 7: Boxplots 35

Appendix 8: QQ Plots showing Normality 36

Appendix 9: Decision Tree Regressor 37

Appendix 10: K-Means Clustering 39

Appendix 11: Customer Reviews Analysis 41

Appendix 12: Top 20 Positive and Negative Reviews and Summaries 44

44

References 46

Business Problem:

Turtle Games, a global games manufacturing and retail company, wants to improve their overall sales performance. They would like their customers' demographics and sales data to be analysed to investigate,

- How do customers accumulate loyalty points
- If their descriptive statistical analysis can be used to create predictive models.
- If customers can be segmented into groups for targeted marketing campaigns to increase sales.
- If customers' feedback (reviews) can be analysed to understand their sentiments for improving product development and business.

To help the marketing department understand the root cause of why the customer data is of low quality, the 5 Why technique was used. (Appendix 1)

Analytical Approach:

Data Wrangling:

To conduct exploratory and predictive analysis, the dataset 'turtle.csv' was imported the into Jupyter notebook and later into RStudio for analysis. Important libraries in Jupyter notebook and RStudio were imported to read, explore, analyse, manipulate, wrangle, and visualise data. The data set was sense checked for missing and duplicate values and outliers as well. These were found to be none. The original dataset had 11 columns and 2000 rows. Irrelevant columns ('language' and 'platform') were removed, and the two columns ('renumeration' and 'spending_score(1-100))' were renamed 'salary' and 'spending' respectively for simplicity. The edited file consisting of 9 columns (Age, Salary, Spending, Summary, Review, Education, Gender, Product and Loyalty) and 2000 rows was saved, and renamed 'clean_reviews.csv', and was used in EDA.

Statistical Analysis:

Statistical analysis was carried out on numerical variables to understand data distribution and correlation analysis was conducted in Python (Jupyter Notebook) and R (RStudio) to explore relationship between the variables.

Visualisations including bar plots, boxplots, scatterplots and histograms were created in RStudio and Jupyter to better understand correlation between variables and if the data was normally distributed. Different tests were run to test for skewness, kurtosis, homoscedasticity and multicollinearity of the variables.

Using Group by() and Aggregate() function, the summary statistics of the whole data set was calculated to investigate average loyalty points earned by customers based on gender, education, salary and spending score (Appendix 5).

Regression Analysis:

Using Python and RStudio, Simple Linear Regression was performed using OLS Method with each predictor variable (salary, spending, age and product) with target variable (Loyalty). The regression plots in all cases showed patterns which suggested the presence of heteroscedasticity. Therefore, logarithmic transformation was applied on the dependent variable ‘loyalty’, and several models were built using multilinear regression as well with and without log transformation, since none of the single predictor variables seemed to affect the accumulation of the loyalty points. Log transformation of Loyalty did improve the R-Squared values, but heteroscedasticity still prevailed. Multicollinearity and prediction errors were checked using VIF and RMSE / MAE metrics calculations.

Next, the decision tree regressor was used to understand how customer features (like salary, spending, age) contribute to loyalty point accumulation. Before creating the model, the importance of features was also conducted to use the relevant variables for X (predictors). Data preparation involved encoding categorical variables and removing irrelevant columns (review and summary). The data was split in 70/30 ratio. The model was fitted onto the training data and its accuracy was tested in terms of the values for MAE and RMSE of the test data. The resulting tree had to be pruned to improve its interpretability and to avoid overfitting.

K-Means Clustering:

To segment customers based on salary and spending, and to guide targeted marketing strategies, customer segmentation was performed using k-means clustering. Scatterplots and pair plots were created to explore potential clusters which showed clear correlation between spending and salary. To evaluate optimal cluster count, the Elbow and Silhouette methods were used. The Elbow method showed a sharp decrease in WSS (Within the Sum of Squares) values up to 5 clusters, then a plateau. Highest silhouette score was also at 5 clusters. The final model was built using 5 clusters.

Sentiment Analysis:

Sentiment analysis was also performed using Natural Language Processing (NLP) techniques and applying libraries like nltk and TextBlob. Data was cleaned (lowercased, punctuation removed, stopwords filtered), tokenised, and visualised through word clouds and frequency distributions. Polarity (sentiment) and subjectivity were computed to quantify emotional tone and objectivity of reviews.

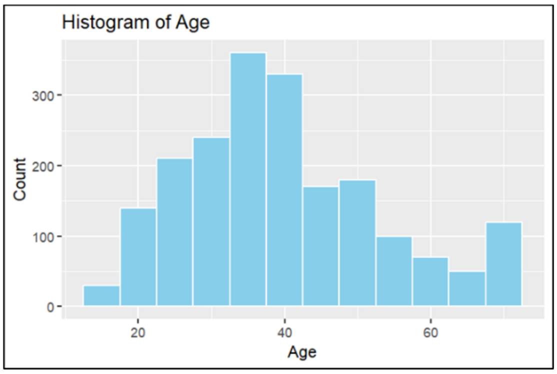
Insights and Recommendation:

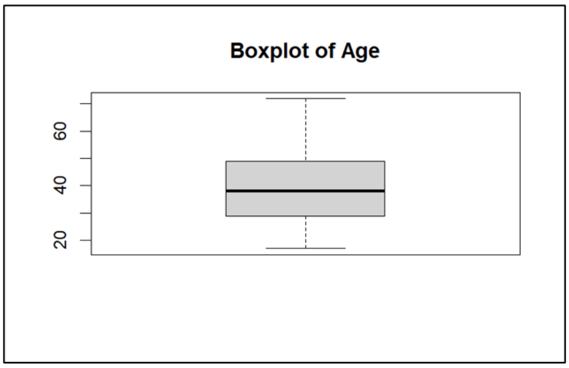
Customer Loyalty Insights

1. Descriptive Analysis

- **Age:** The histogram and boxplot show that customer age is right skewed with most customers clustered around the 30s and early 40s. The median age is approximately 38. There are no significant outliers.

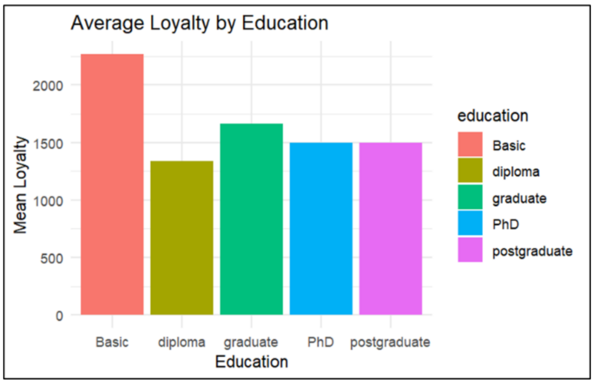
- **Salary:** The distribution is approximately symmetric, with values ranging from around 12 to 112. The median salary is about 47. A few higher-end salaries stretch the upper bound but no extreme outliers.
- **Spending Score:** This variable is more uniformly spread from 1 to 100, with some spikes in the 20s, 50s, and 70s. The boxplot confirms there is no major skewness.
- **Loyalty Points:** Loyalty is heavily right skewed with some very high values (max 6847). The median loyalty is 1276, and many outliers appear in the upper range.
- **Product:** The boxplot shows a wide range of product codes with a roughly symmetric distribution and no major outliers, suggesting a consistent spread of product types in the dataset. The histogram reveals that some product codes appear far more frequently than others, indicating that certain items are much more popular or commonly sold, while others are less frequent in customer interactions (Appendix 4).

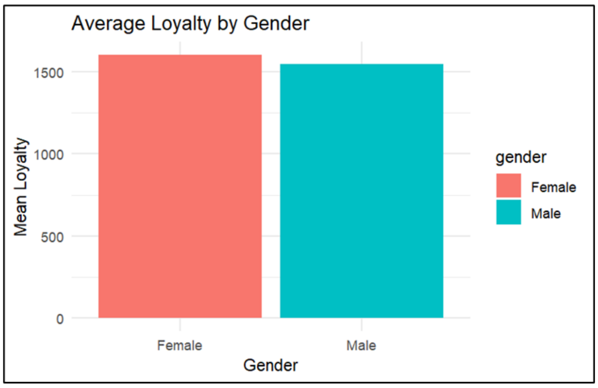




2. Grouped Loyalty Analysis: (Appendix 5)

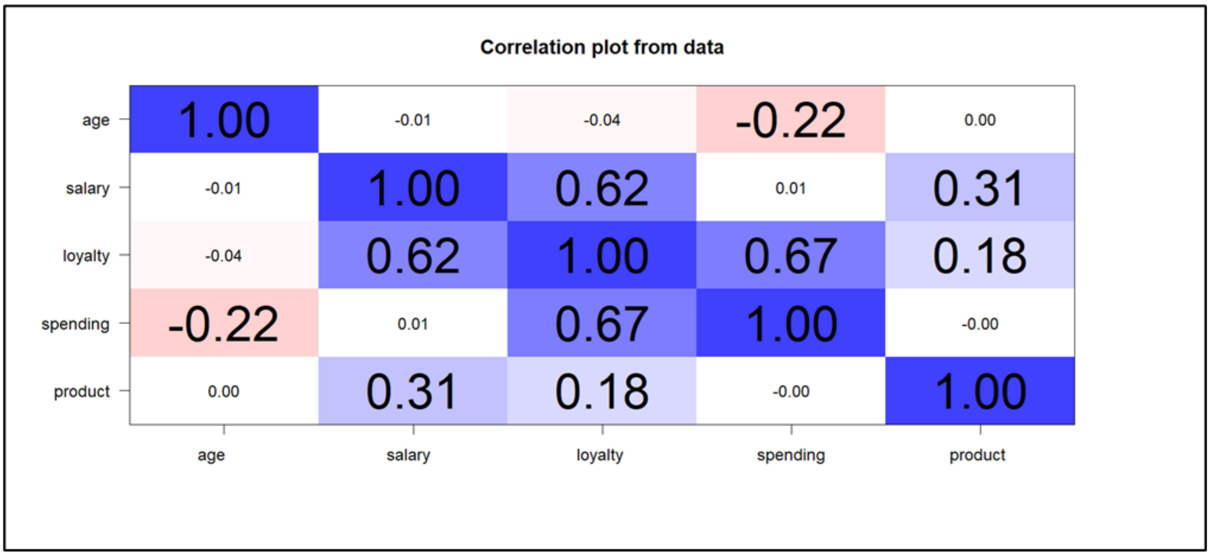
- **By Gender:** Female customers have a slightly higher average loyalty (1601 vs 1549 for males). Females also show slightly less variance in loyalty points. The gender distribution of the data set reveals that there are more female customers than the males.
- **By Age:** Loyalty peaks for customers in their early 30s (32-34), suggesting this age group might be more engaged or targeted. Loyalty then declines gradually with age.
- **By Education:** Customers with Basic education surprisingly have the highest average loyalty. Loyalty decreases as education level increases. This might suggest differing engagement levels by demographic (Appendix 7).
- **By Salary:** Loyalty rises steadily with salary, especially after a salary level of 60. This positive trend highlights that higher-income customers are likely more loyal.
- **By Spending:** A strong upward trend is observed—as spending score increases, so does average loyalty, reinforcing the idea of engagement and value alignment.
- **Spending & Salary Combined:** A bubble plot confirms the strongest loyalty occurs in customers with high salary and high spending scores.





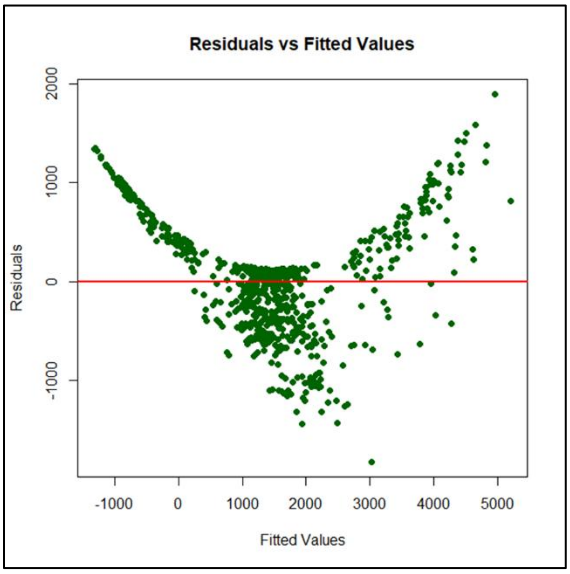
3. Correlation Analysis: (Appendix 3)

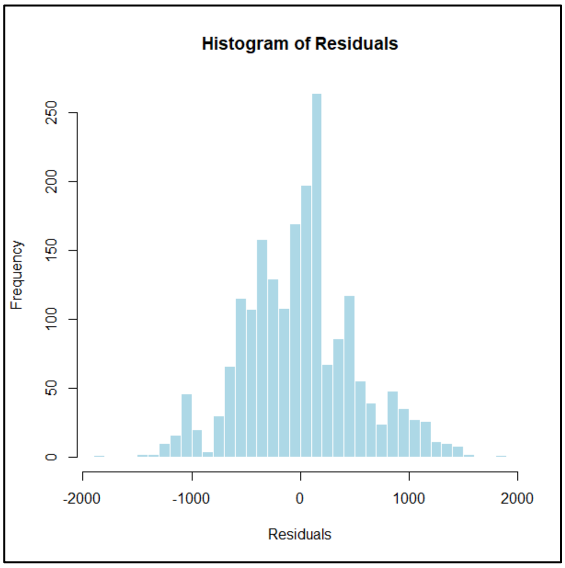
The analysis shows that product engagement has a moderate positive relationship with salary (0.31), suggesting that wealthier customers tend to interact more with products**.** Loyalty is strongly tied to both spending (0.67) and salary (0.62)**,** while its link to product is weaker (0.18). Spending and age show no meaningful connection to product, implying product preferences are more influenced by income than by age or spending patterns. This is also noticed in Scatterplots (Appendix 6).



4. Regression Models Overview

Spending and Salary are the strongest predictors of Loyalty. Of the seven linear regression models that were evaluated to predict loyalty in RStudio, Model 2 (Salary + Age + Spending) proved the most suitable model for predicting customer loyalty. It balances accuracy, simplicity, and interpretability (RMSE = 513.31, MAE = 394.98). Model 2's residuals were normally distributed, supporting the assumption of normality. All VIF values were under 1.1, indicating no multicollinearity among predictors. However, the residuals vs fitted plot showed a U-shape, suggesting some non-linearity. Residual plots also indicate non-constant variance, suggesting heteroscedasticity. Nonetheless, predictions are more accurate than other models. See regression summaries tables of different models in Appendix 2.





Model	Predictors	RMSE	MAE	Notes
Model 1	Salary, Age, Spending, Product	513.30	394.96	Product not significant
Model 2	Salary, Age, Spending	513.31	394.98	Best model
Model 3	Salary, Spending	533.74	414.83	Slightly less accurate
Model 4	log(Loyalty) ~ Salary + Age + Spending	905.99	485.92	Poorer fit
Model 5	log(Loyalty) ~ Salary + Spending	866.11	501.40	Slightly better than Model 4
Model 6	Salary only	1010.55	716.30	Weakest model
Model 7	Spending only	949.71	668.52	Stronger than salary-only model

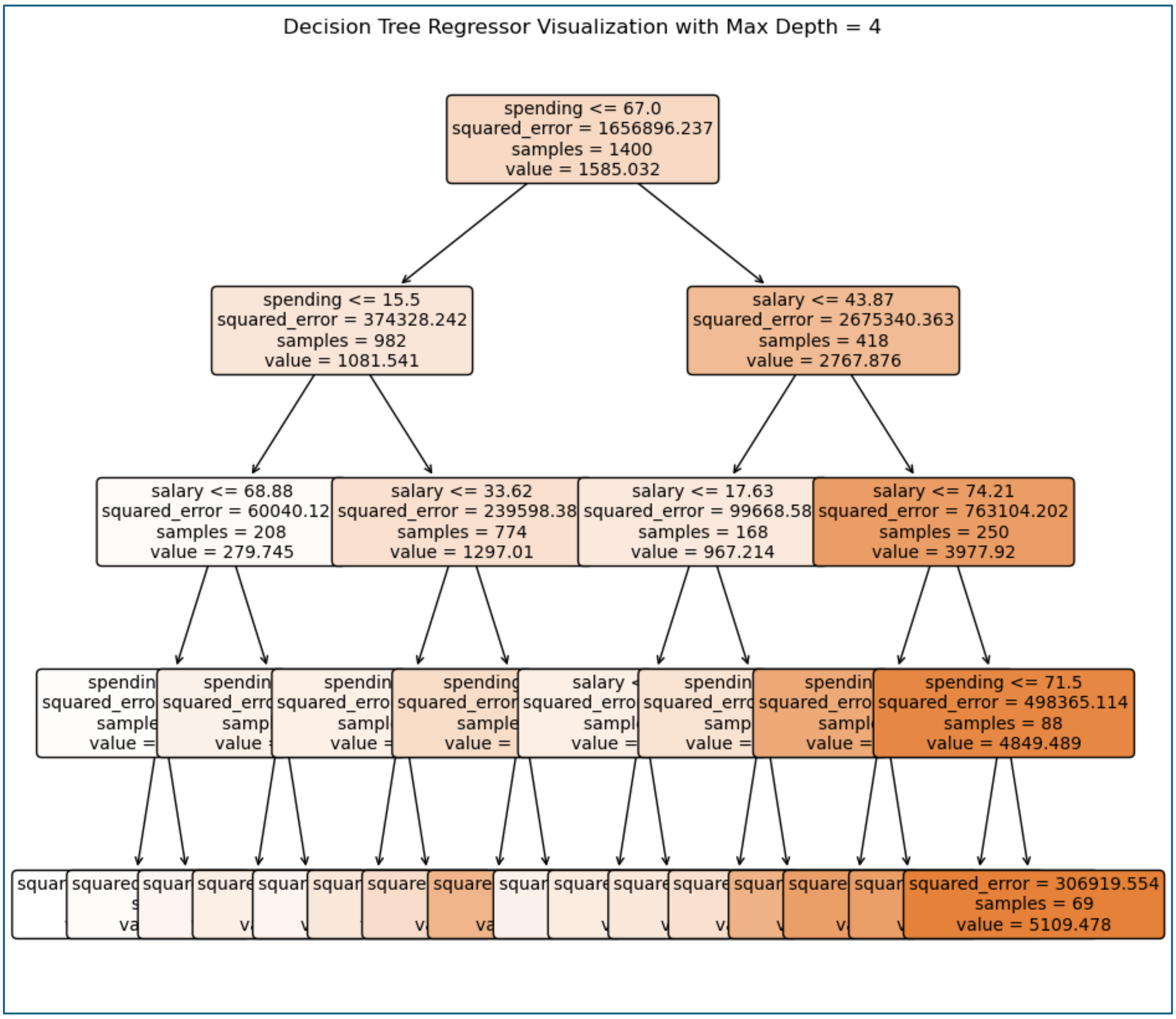
Model comparison based on RMSE and MAE

5. Decision Tree Regressor: (Appendix 9)

The decision tree regressor models showed that salary and spending are the biggest drivers of customer loyalty followed by age. The tree with the maximum depth of 3 seemed like a better model in terms of interpretability and simplicity but with the added risk of increasing mean absolute error and potential underfitting. In other words, it may not capture enough details or trends in the data. There was no change in the R-Squared value (0.9961). However, the value of MAE increased by pruning.

Model Iterations:

- **Model 1**: Used all variables except review and summary. It showed high R^2 (0.9938) but also high RMSE (100.16), indicating possible overfitting.
- **Model 2**: Focused on the top two predictors (spending and salary). R^2 decreased to 0.9839, and RMSE increased to 161, showing reduced overfitting but lower accuracy.
- **Model 3**: Added age to the predictors. It produced the best performance: $R^2 = 0.9961$ and MAE = 26, making it the most accurate and preferred model.



6. K-Means Clustering: (Appendix 10)

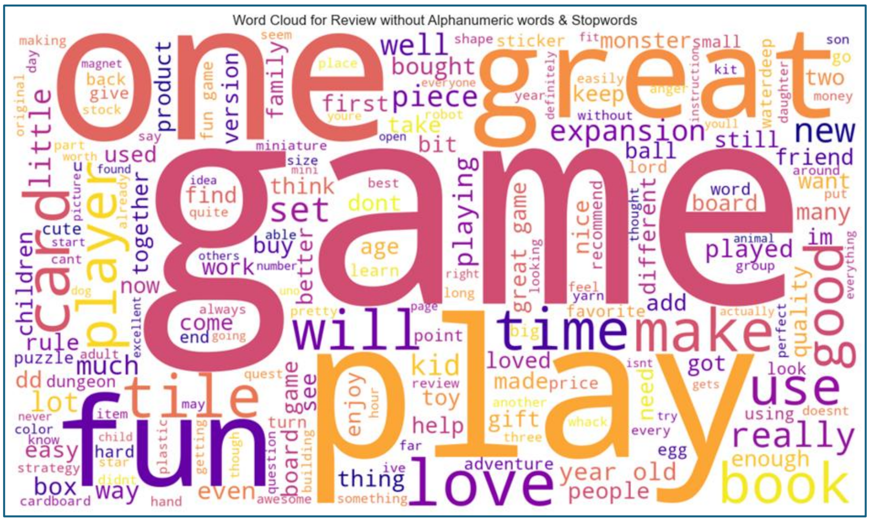
Customer segmentation with k-means clustering yielded 5 Clusters: Customers with moderate income (30-60K£) and moderate spending score (30 - 60). Customers with high income bracket (60-100K£) and low spending score (10-40). Customers earning high income (60-100K£) with high spending score (Over 100). Low earners (20-40K£) with high spending score (60 -100). Cluster 4 belongs to customers with low income (20-40K£) and low spending score (Below 40).

Cluster	Income Range	Spending Score	Description
0 (Red)	£30k–£60k	30–60	Moderate earners/spenders (average shoppers). They represent average shoppers who like to spend within their means.
1 (Blue)	£60k–£100k	10–40	High income, low spending. These customers should be engaged with offers.
2 (Green)	£60k–£100k	>100	High earners & high spenders — ideal for premium products and offers.
3 (Purple)	£20k–£40k	60–100	Low income, high spenders — loyal but price-sensitive
4 (Yellow)	£20k–£40k	<40	Low income, low spending — These customers are restricted by budget limitations.

6. Sentiment Analysis of Customers reviews: (Appendix 11)

Most frequently used positive words included 'fun, game, play, great, stars, five, love, excellent'. Top positive reviews scored a sentiment polarity of +1, with phrases like 'awesome book', 'wonderful product', and 'perfect condition'. Top negative reviews had polarity scores close to -1, with terms describing frustration or disappointment like 'incomplete kit', 'boring' and 'difficult to use'

Overall, the frequency distribution graphs, word clouds and sentiment polarity scores indicate that most reviews were neutral to slightly positive, suggesting general satisfaction. There were very few extreme positive or negative reviews showing a balanced and moderate customer experience. Neutral reviews may indicate that customers are satisfied but not enthusiastic probably because the products meet their expectations but do not exceed them. The presence of positive sentiments mean that most customers are happy and satisfied.



Word Cloud for Customer ‘Reviews’ column

7. Recommendations:

Boost Loyalty with Targeted Offers

- To increase customer retention and loyalty, Turtle Games marketing team could use customer segmentation information in increasing their sales by targeting the different types of customers. Customers in cluster 1 (high earners, low spending) can be offered incentives and promotions to increase their spending score. Those in cluster 2 (high earners, high spending) can be targeted with premium offers to attract more custom. Similarly, customers with moderate spending and low/ moderate incomes could be engaged with loyalty programs or offered products discounts.

- Personalise Offers for customers with basic education who show high loyalty despite potentially lower income levels.

Refine Customer Segmentation

- Focus loyalty campaigns on ages 32–34, the most engaged group.
- Use salary and spending to tailor marketing strategies for each cluster.

Improve Product & Review Experience

- Highlight top-performing products in promotions and investigate less frequent product codes for possible redesign or bundling.

- The marketing team could shift the neutral sentiment of the less enthusiastic customers to a positive one by focussing on improving their customer experience and addressing common pain points. They could do this by encouraging detailed reviews using structured survey questions to gather information about customers' experience with their products and service. The concern about low quality data can be addressed by encouraging customers to leave detailed reviews. This can be achieved by asking specific questions during the review process, such as asking for comments on product

quality, delivery, and customer service separately. They could then use this knowledge to understand what features click with the customers. This information can be used to improve product quality and customer service and marketing strategies.
- The ‘Product’ column contained product codes, not purchase counts, limiting its value. Future datasets should track purchase frequency per product to better understand demand.

Leverage Insights from Text Analytics

- Act on common positive themes (fun, great, love) in marketing.
- Address usability concerns from negative reviews to reduce dissatisfaction.

Enhance Data Strategy

- Invest further in text analytics and customer feedback mining.
- Enhance review collection processes to improve data quality and volume.

Appendix 1: The 5 Whys Technique

The 5 Why Technique:

1. Why is the customer reviews data of low quality?

- Because many reviews are incomplete or vague.

2. Why are the reviews incomplete or vague?

- Because customers are not providing enough detail when submitting their reviews.

3. Why are customers not providing enough detail?

- Because the review submission process may not ask for detailed feedback or specific information.

4. Why is the review submission process not asking for detailed feedback?

- Because the review form may be too simple, lacking prompts or guidelines to encourage more detailed responses.

5. Why is the review form too simple and lacks prompts?

- Because the company may not have invested in improving the review system or doesn't prioritize collecting in-depth feedback.

Root Cause:

The review submission process is inadequately designed, lacking structured prompts or questions that guide customers to provide detailed and high-quality feedback.

Solution Ideas Based on Root Cause:

- **Redesign the review form** to include specific questions or prompts that encourage detailed feedback (e.g., asking about product features, customer service experience, etc.).
- **Incentivize detailed reviews** by offering discounts or loyalty points for thorough reviews.
- **Implement review moderation** or use text analytics to filter and categorize useful reviews while flagging low-quality ones.

Appendix 2: Regression Summaries Tables

Figure 2.1: Predicting Loyalty with Age, Product, Salary and Spending.

- Spending and salary are the strongest predictors of loyalty.
- Age also contributes positively, but less so.
- Product has no significant effect ($p = 0.75$), meaning it doesn't help explain loyalty variation in this model.

Regression Summary: Predicting Loyalty with Age, Product, Spending and Salary				
term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	-2200.255	53.108	-41.430	0.00
salary	34.059	0.522	65.242	0.00
age	11.062	0.869	12.729	0.00
spending	34.183	0.452	75.620	0.00
product	-0.001	0.004	-0.319	0.75

Figure 2.2: Predicting Loyalty with Age, Salary and Spending.

- This model shows a strong, positive, and statistically significant relationship between all three predictors and loyalty.
- All p-values are essentially zero — meaning these predictors are highly significant.
-

Regression Summary: Predicting Loyalty with Age, Salary & Spending

term	estimate	std.error	statistic	p.value
(Intercept)	-2203.060	52.361	-42.075	0
salary	34.008	0.497	68.427	0
age	11.061	0.869	12.730	0
spending	34.183	0.452	75.638	0

Figure 2.3: Predicting Loyalty with Salary and Spending.

Regression Summary: Predicting Loyalty with Salary & Spending				
term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	-1700.305	35.740	-47.575	0
salary	33.979	0.517	65.769	0
spending	32.893	0.458	71.845	0

Figure 2.4: Predicting loyalty points with Salary (Renumeration)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                loyalty   R-squared:                       0.380
Model:                            OLS   Adj. R-squared:                  0.379
Method:                 Least Squares   F-statistic:                     1222.
Date:                Fri, 27 Sep 2024   Prob (F-statistic):          2.43e-209
Time:                        00:47:46   Log-Likelihood:                -16674.
No. Observations:                2000   AIC:                         3.335e+04
Df Residuals:                    1998   BIC:                         3.336e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -65.6865     52.171     -1.259      0.208    -168.001      36.628
salary        34.1878      0.978     34.960      0.000      32.270      36.106
==============================================================================
Omnibus:                       21.285   Durbin-Watson:                   3.622
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.715
Skew:                           0.089   Prob(JB):                     1.30e-07
Kurtosis:                       3.590   Cond. No.                         123.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

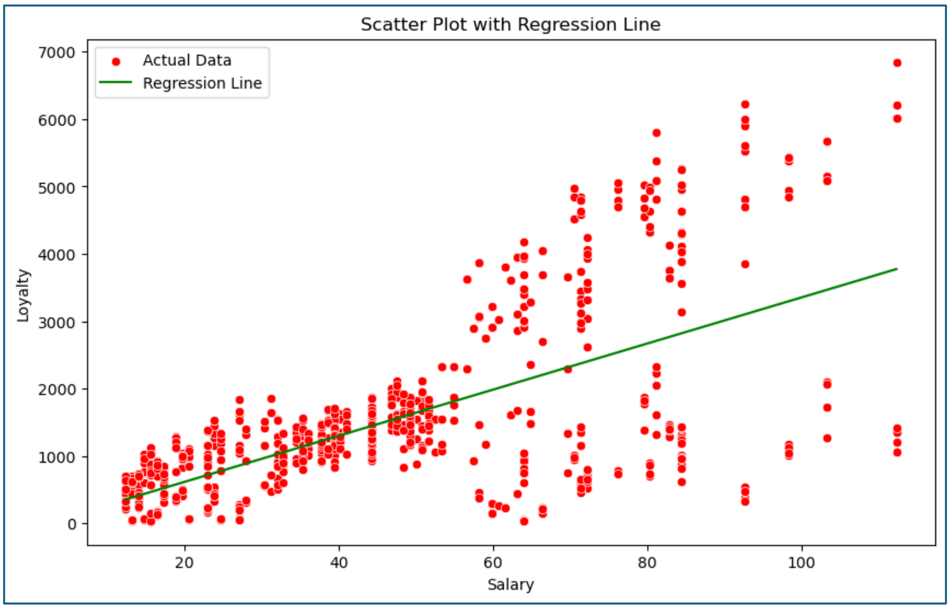
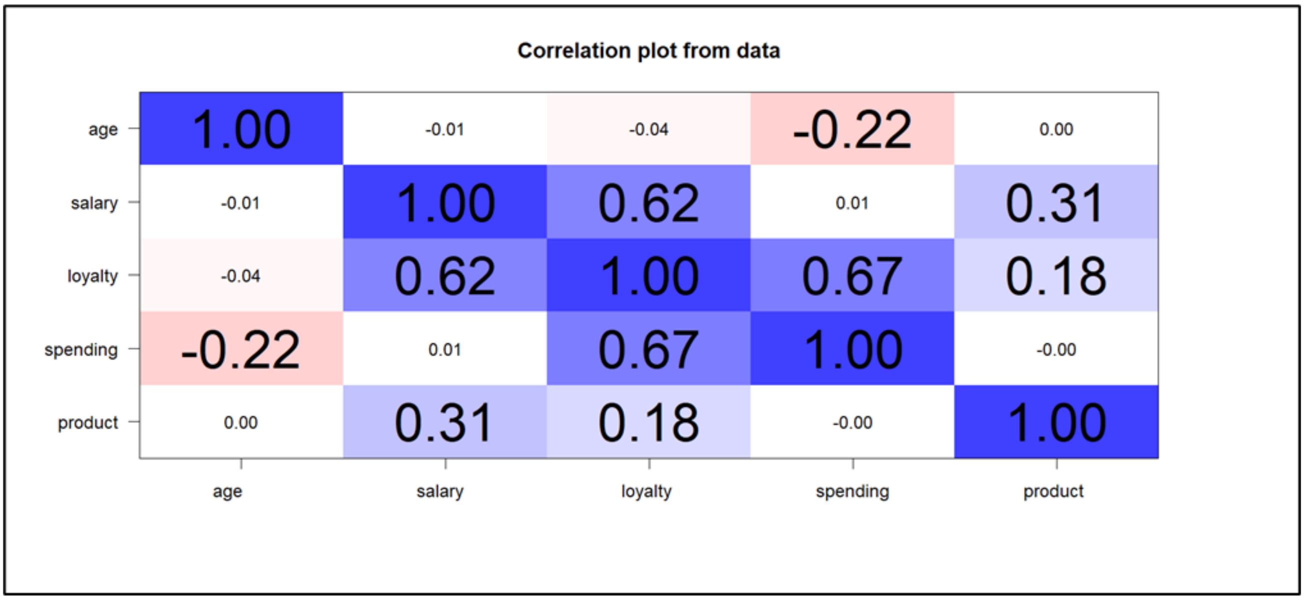


Figure shows the uneven variance of data points along the line of regression shows the presence of heteroscedasticity.

Figure 2.5: Predicting loyalty points log loyalty and Spending.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  log_y   R-squared:                       0.519
Model:                            OLS   Adj. R-squared:                  0.518
Method:                 Least Squares   F-statistic:                     2153.
Date:                Fri, 27 Sep 2024   Prob (F-statistic):         1.44e-319
Time:                        00:47:52   Log-Likelihood:                -2146.7
No. Observations:                2000   AIC:                             4297.
Df Residuals:                    1998   BIC:                             4309.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.5740      0.034    162.833      0.000       5.507       5.641
spending       0.0282      0.001     46.400      0.000       0.027       0.029
==============================================================================
Omnibus:                      247.764   Durbin-Watson:                   0.562
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              344.804
Skew:                          -1.000   Prob(JB):                     1.34e-75
Kurtosis:                       3.366   Cond. No.                         122.
==============================================================================
```

Appendix 3: Correlation Matrix

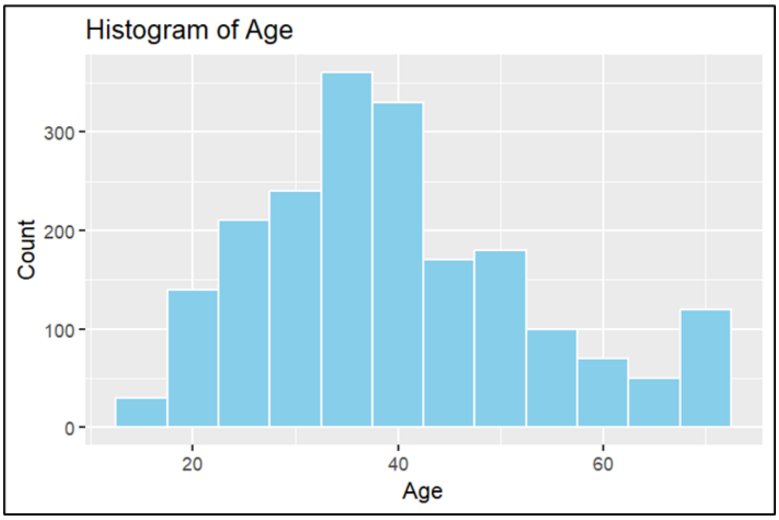


The analysis shows that **product engagement has a moderate positive relationship with salary (0.31)**, suggesting that **wealthier customers tend to interact more with products**. Loyalty is strongly tied to both **spending (0.67)** and **salary (0.62)**, while its link to product is weaker (0.18). **Spending and age show no meaningful connection to product**, implying product preferences are more influenced by income than by age or spending patterns.

Appendix 4: Histograms, Boxplots

Figure 4.1: Age Distribution

- **Histogram**: Age distribution is slightly right-skewed, with a concentration between ages 30–45. This indicates a younger-to-middle-aged customer base.
- **Boxplot**: No extreme outliers are present. Median age is around the mid-30s, with a reasonable spread from early 20s to 60s.



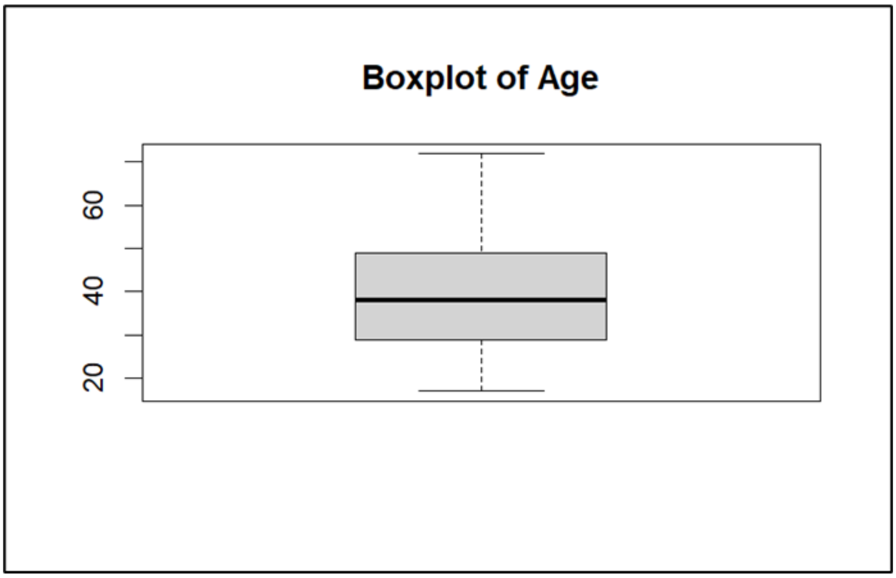
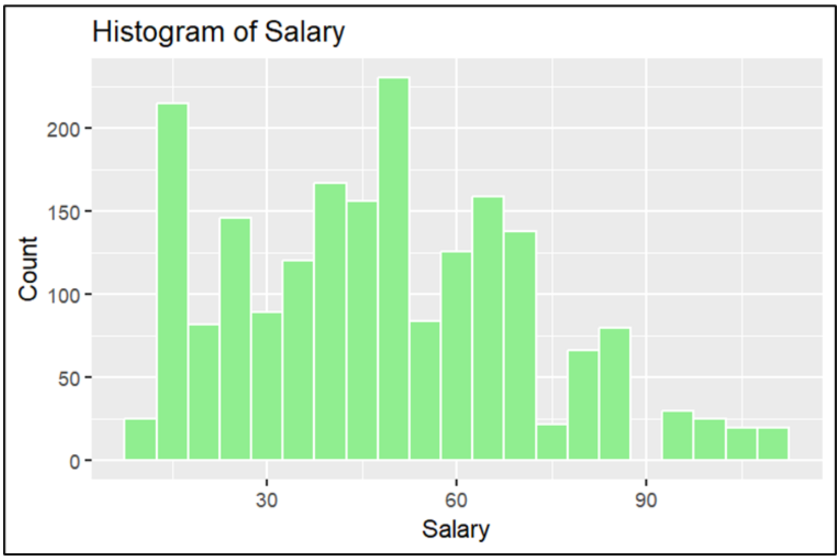


Figure 4.2: Salary Distribution

- **Histogram**: Salary is moderately right skewed, with most individuals earning between 25 to 65 units. A few higher earners extend the distribution's tail.
- **Boxplot**: The interquartile range spans from ~30 to ~70, with some high-end outliers. Suggests a moderately diverse income distribution.



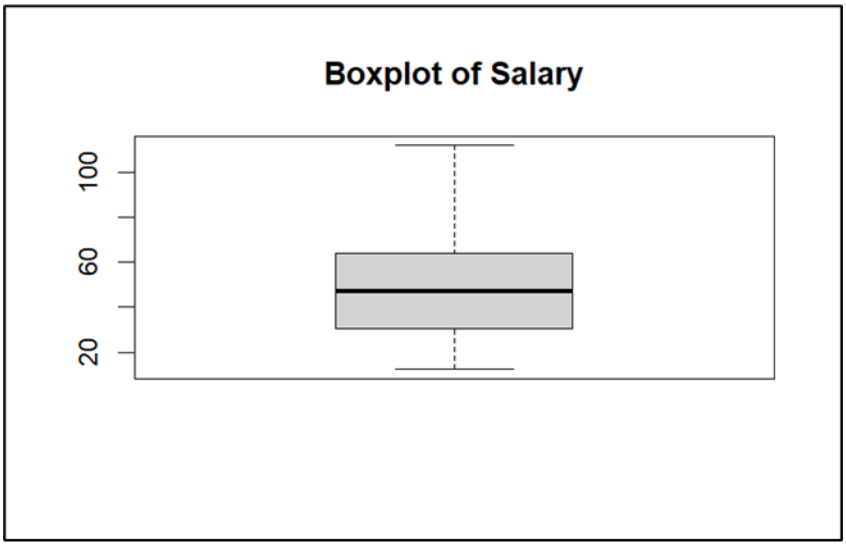
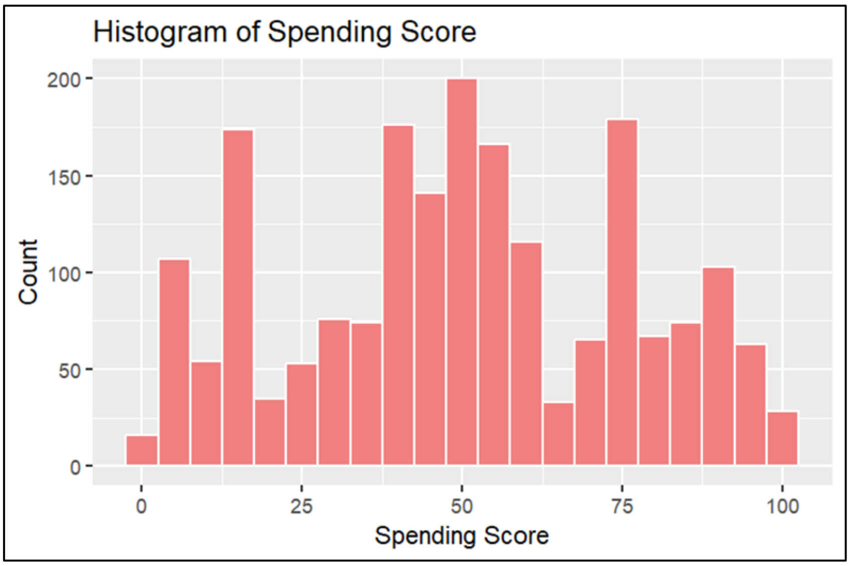


Figure 4.3: Spending Score Distribution

- **Histogram**: Appears bimodal or multimodal with spikes at different score ranges (e.g., 0–25, 50–75). This may indicate distinct consumer groups with low vs high spending.
- **Boxplot**: Even distribution across the range, from 0 to 100. Median and quartiles suggest broad variability with potential outliers on both ends



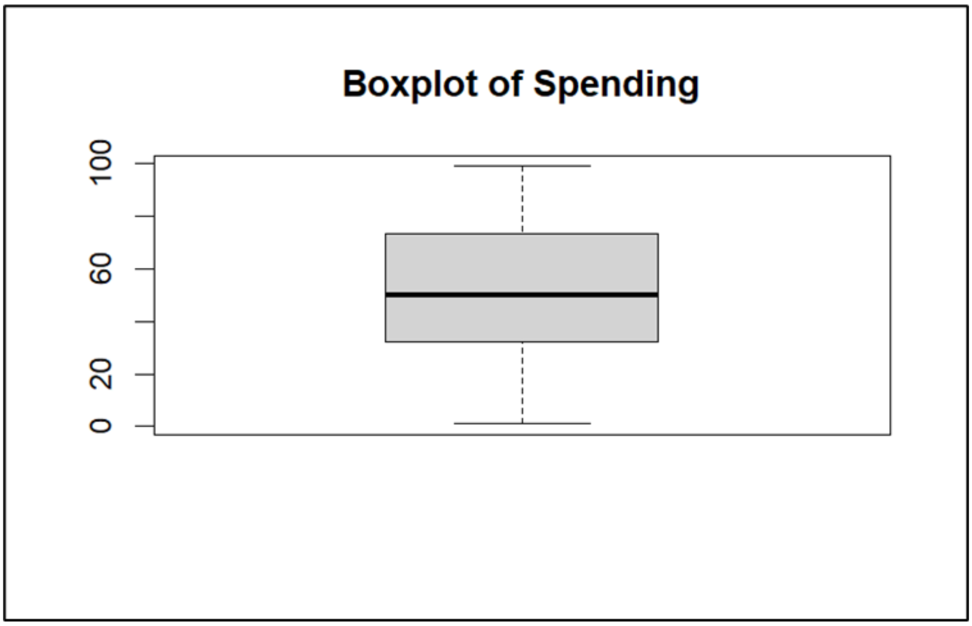
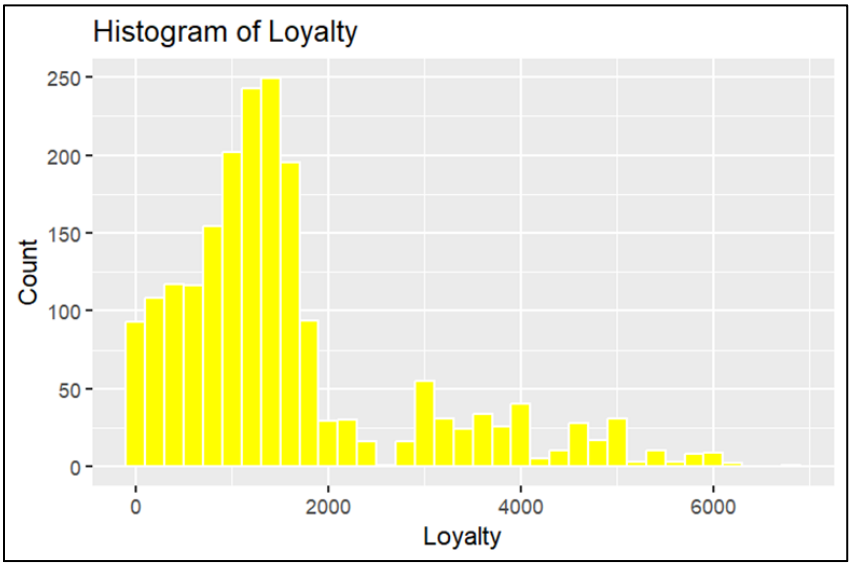


Figure 4.4: Loyalty Points Distribution

- **Histogram**: Strong right-skew with most customers having loyalty points below 2000. A few individuals have extremely high points (>6000), which may be anomalies or highly engaged users.
- **Boxplot**: Significant number of outliers above the upper quartile. Indicates most customers earn modest loyalty points, but a minority accumulate very high values.



mean loyalty	median loyalty	Maximum loyalty	Minimum loyalty	Standard Deviation Loyalty
1578.032	1276	6847	25	1283.24

Summary Statistics of Loyalty Points for all customers

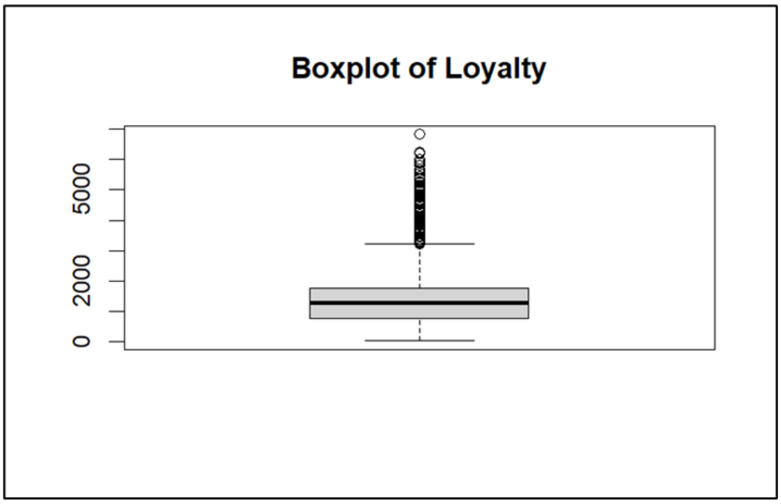
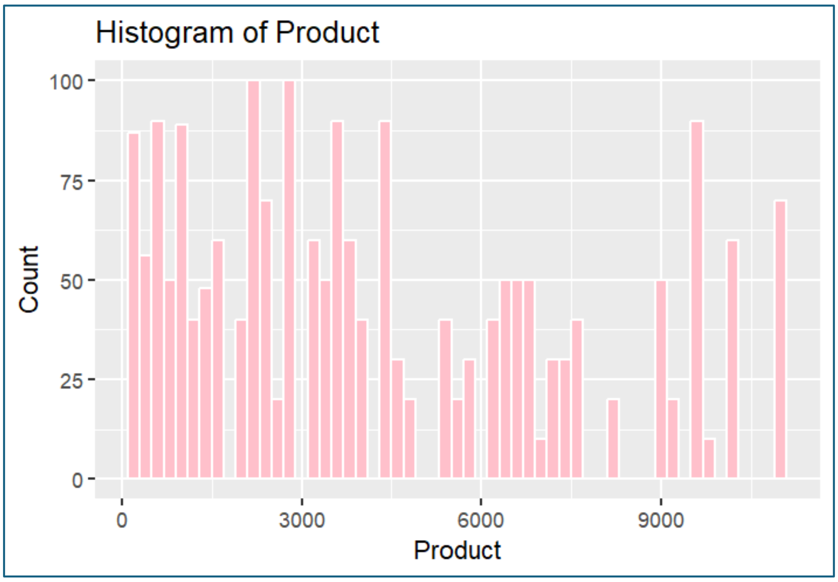


Figure 4.5: Product Distribution

• The **boxplot** shows a wide range of product codes with a roughly symmetric distribution and no major outliers, suggesting a consistent spread of product types in the dataset.

• The **histogram** reveals that some product codes appear far more frequently than others, indicating that certain items are much more popular or commonly sold, while others are less frequent in customer interactions. This points to varying popularity among the product offerings.



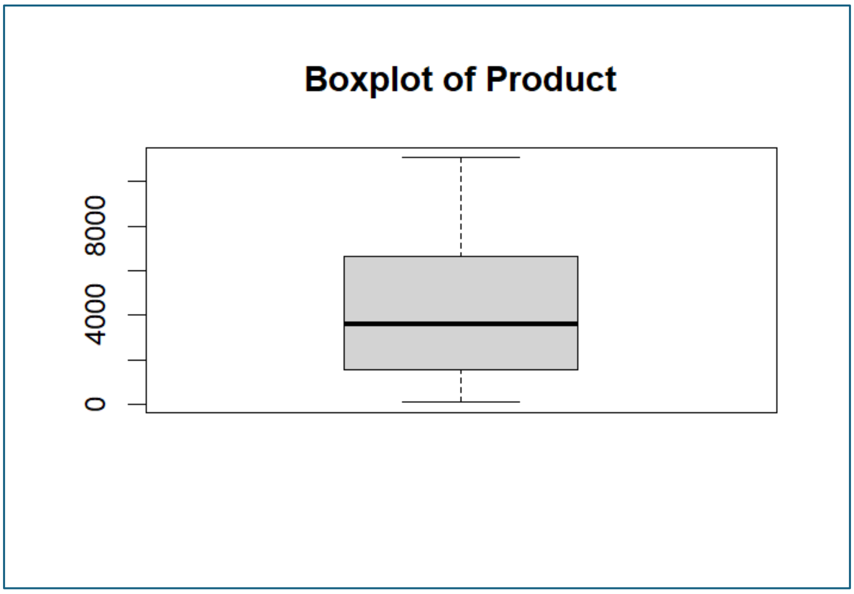
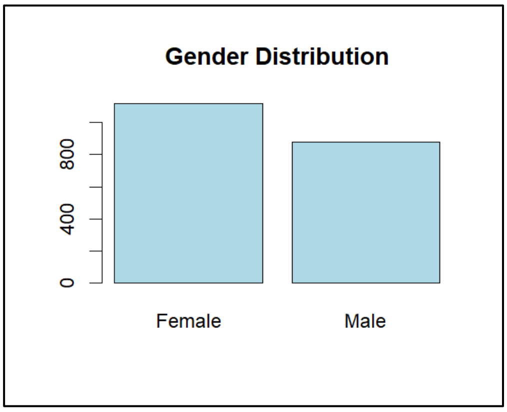


Figure 4.**6: Combined Histogram Facets**

- **Faceted View**: This visualization compares distributions side by side and emphasizes key contrasts:
  - **Age and salary** have a semi-normal to right-skewed shape.
  - **Spending score** shows multimodal behaviour.
  - **Loyalty** is highly right skewed with many low values and long tails.

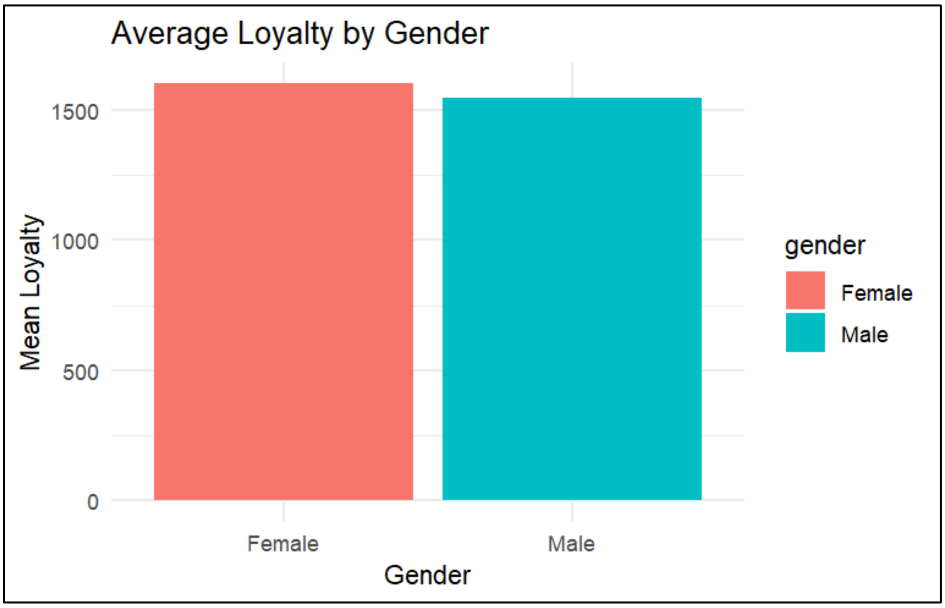
These histograms confirm that none of the variables follow a perfect normal distribution. The loyalty and product variables are positively skewed, while age and salary are more symmetric but not normally distributed. This has implications for the assumptions of linear regression and supports the use of transformations.

Figure 4.7: Gender Distribution



Appendix 5: Average Loyalty Points earned by Customers based on Age, Education, Gender, Spending, Product and Salary.

Figure 5.1: Average Loyalty Points by Gender



Statistical Summary by Gender: On average, females show slightly higher loyalty scores (Mean: 1601) compared to males (Mean: 1549), though the difference is small. The spread (standard deviation) is higher for males (1323 vs. 1251), indicating more variability in male customer loyalty. Both genders have similar medians and minimums, but females reach a higher maximum loyalty score (6847 vs. 6208), suggesting that the most loyal customers tend to be female.

Gender	Mean Loyalty	Median Loyalty	Maximum Loyalty	Minimum Loyalty	Standard Deviation Loyalty
Female	1601	1281	6847	30	1251
Male	1549	1248	6208	25	1323

Figure 5.2: Average Loyalty Points by Education.

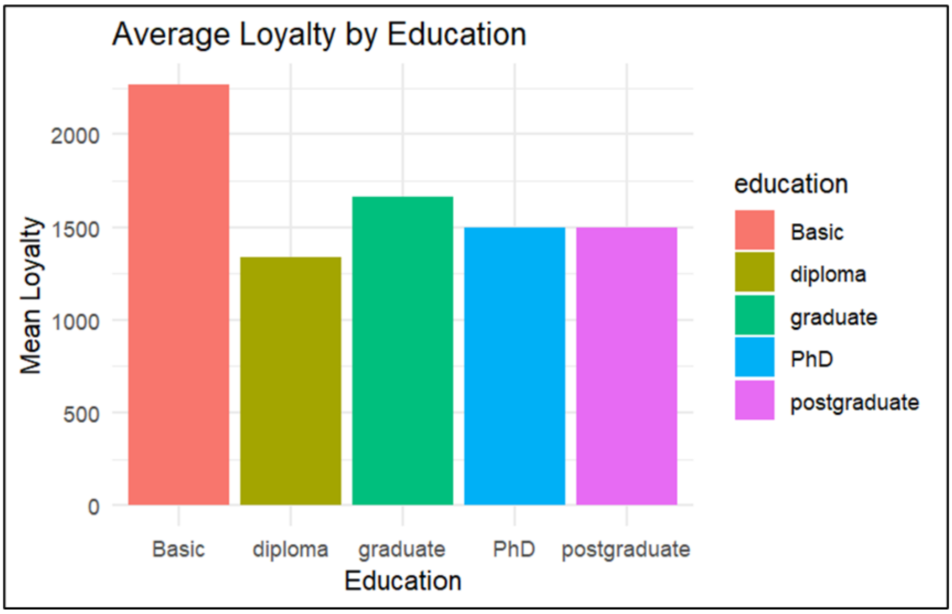


Figure 5.3: Average Loyalty Points by Age.

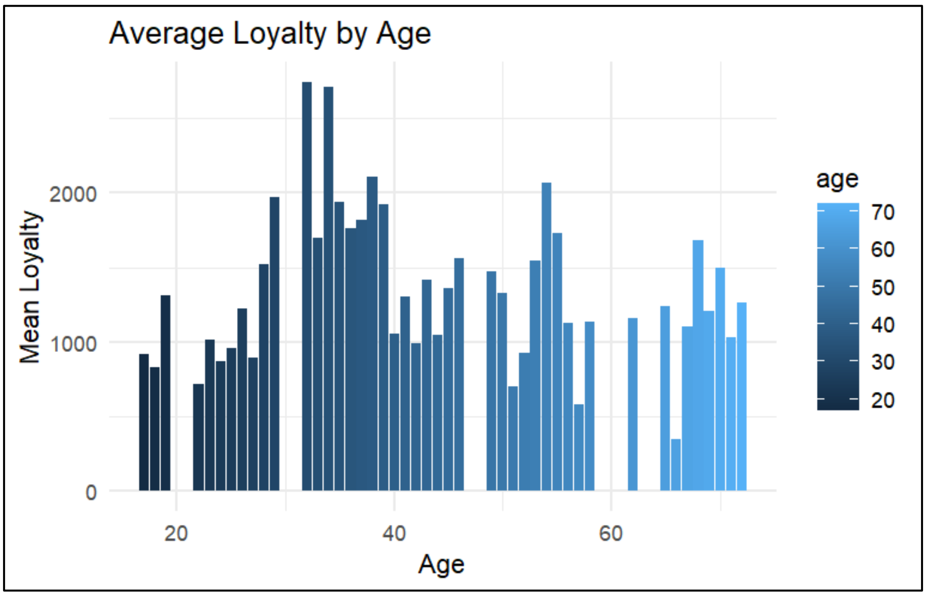


Figure 5.4: Average Loyalty Points by Salary

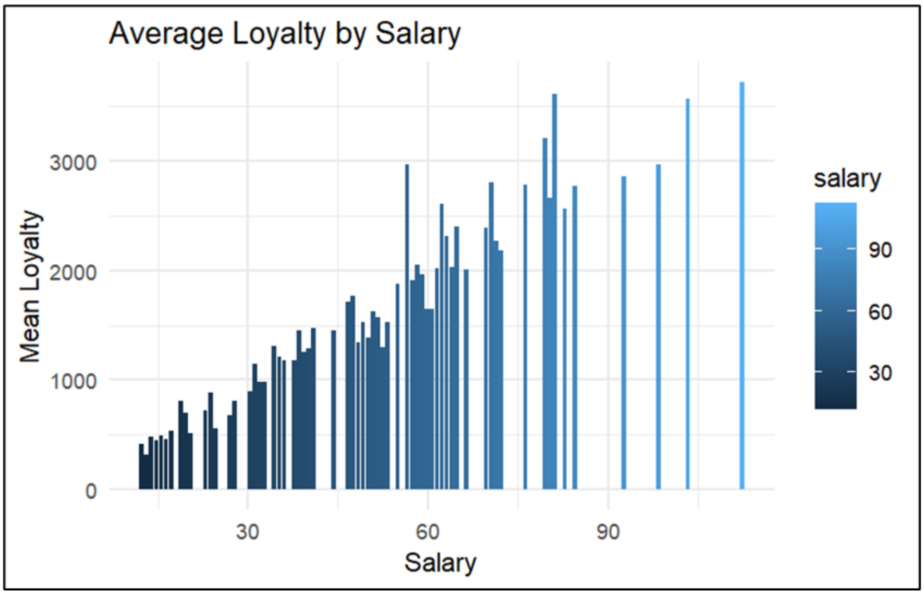


Figure 5.5: Average Loyalty Points by Spending

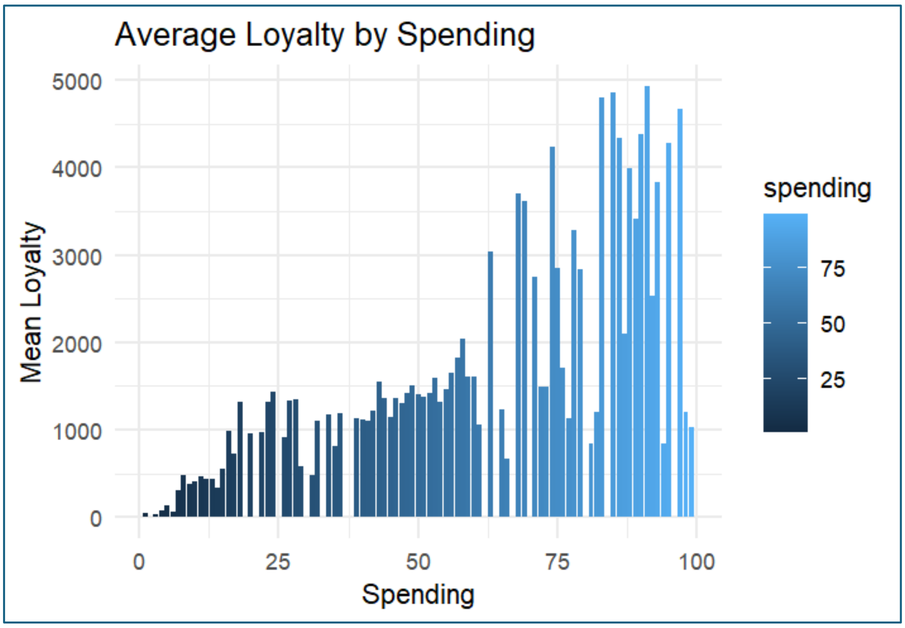


Figure 5.6: Average Loyalty Points by Product

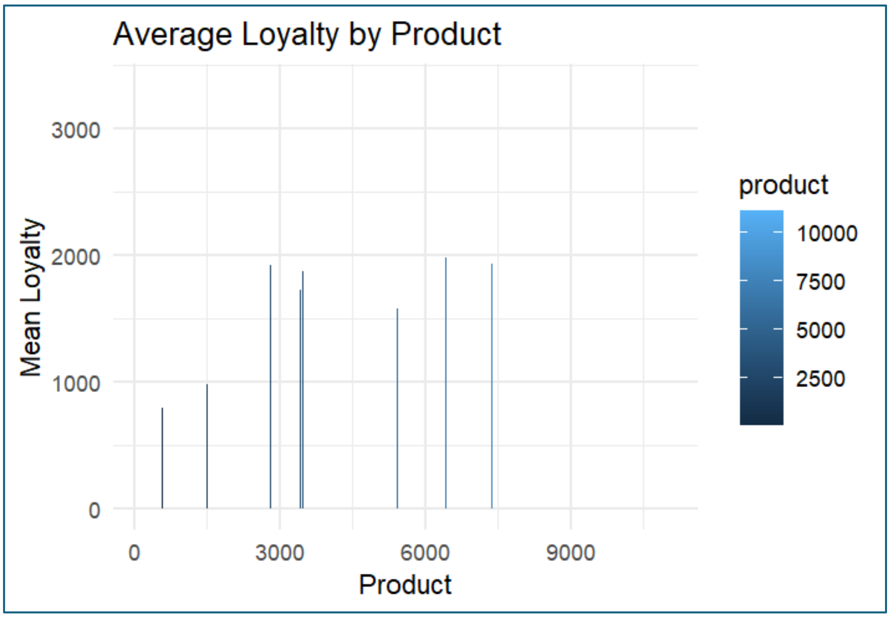
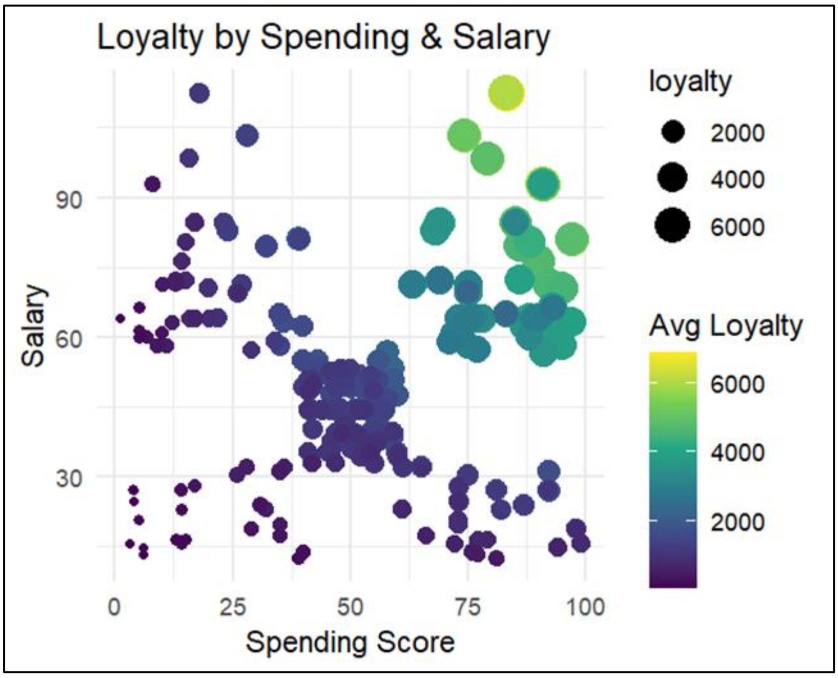


Figure 5.7: Average Loyalty Points by Spending and Salary



Appendix 6: Scatterplots showing Correlation between different Predictor variables (Age, Salary, Spending, Product) and Target variable (Loyalty).

Figure 6.1: Scatterplot showing correlation between Spending and Loyalty

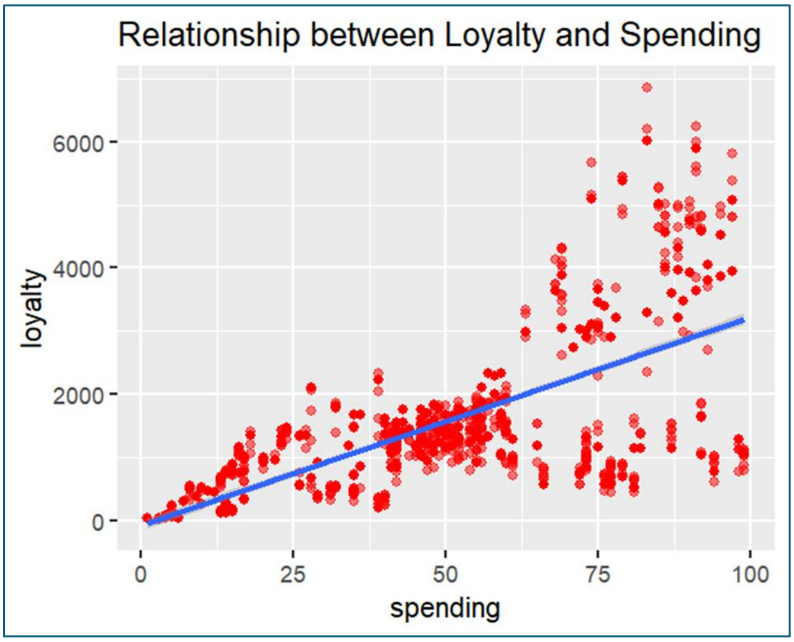


Figure 6.2: Scatterplot showing correlation between Age and Loyalty

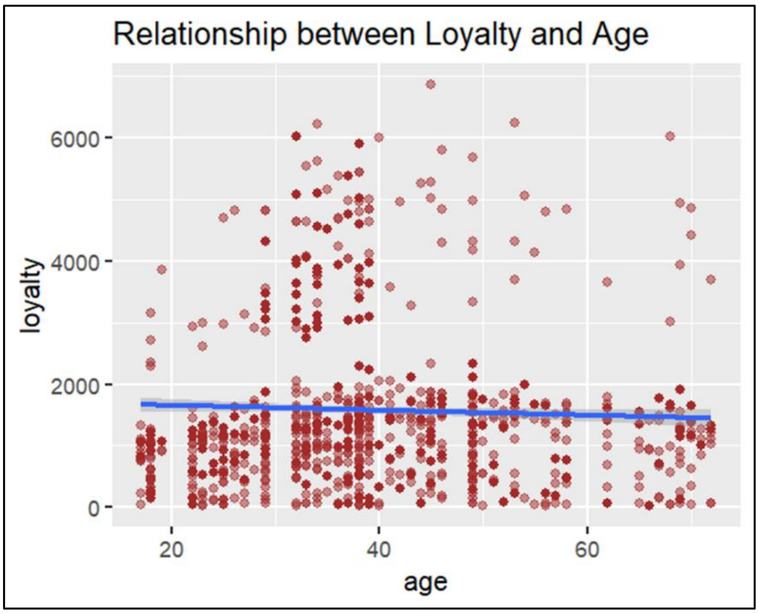


Figure 6.3: Scatterplot showing correlation between Salary and Loyalty

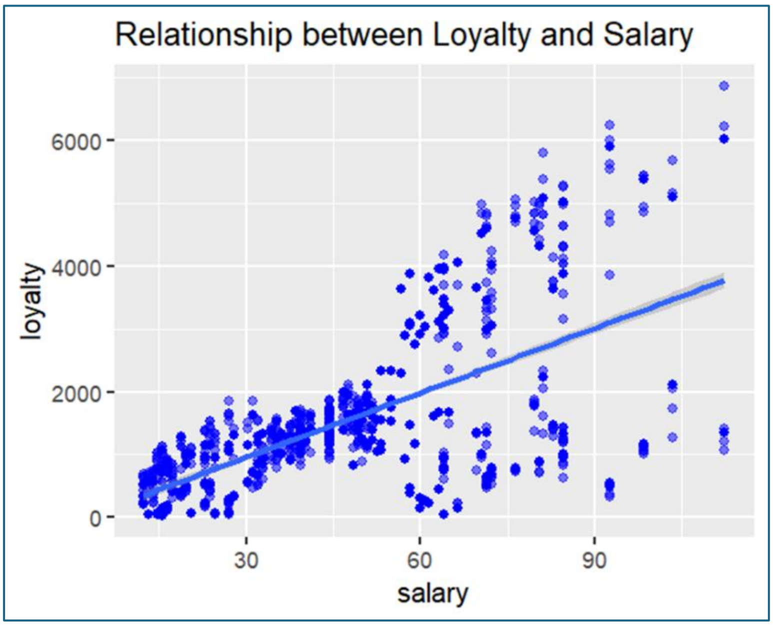


Figure 6.4: Scatterplot between Product and Loyalty

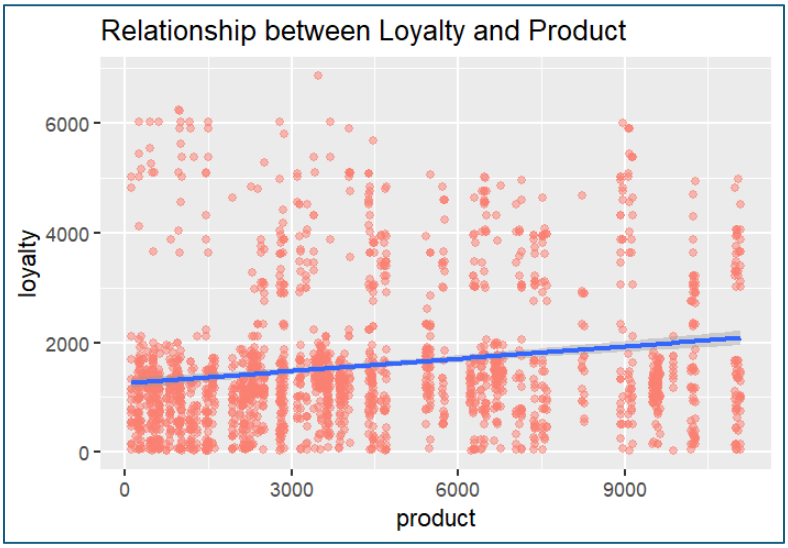
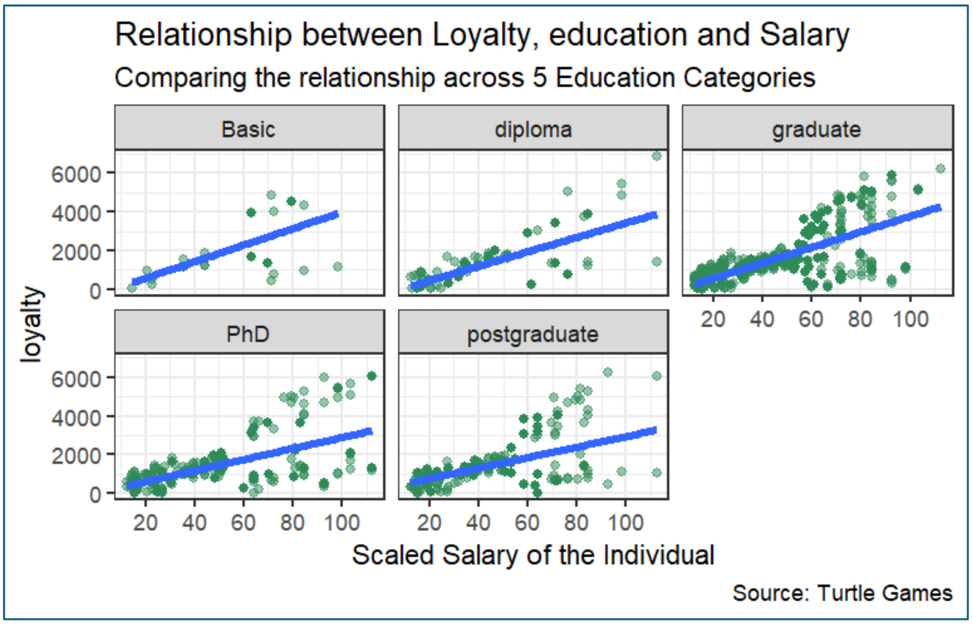
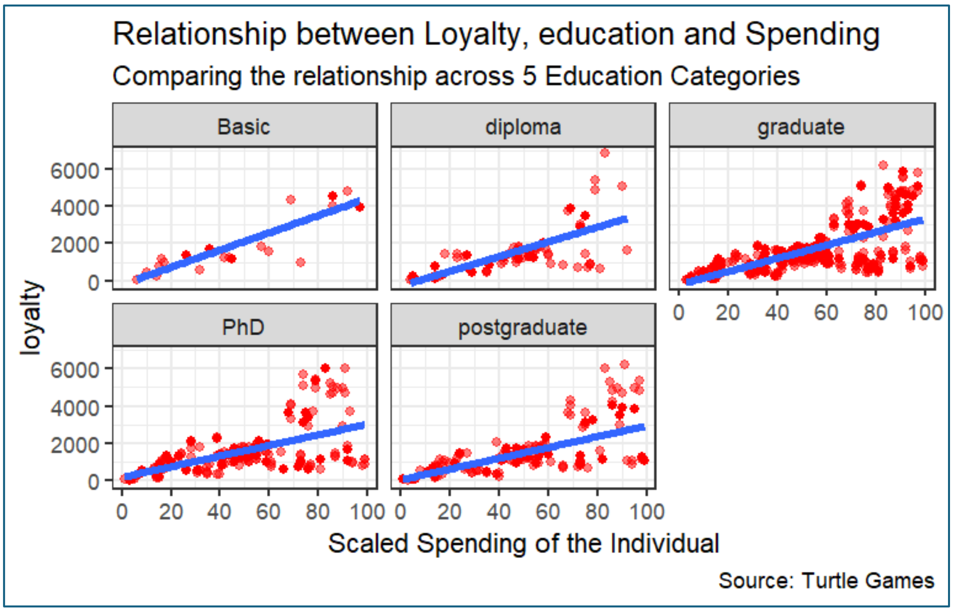


Figure 6.5: Scatterplots showing relationship between Salary, Education and Loyalty Points.



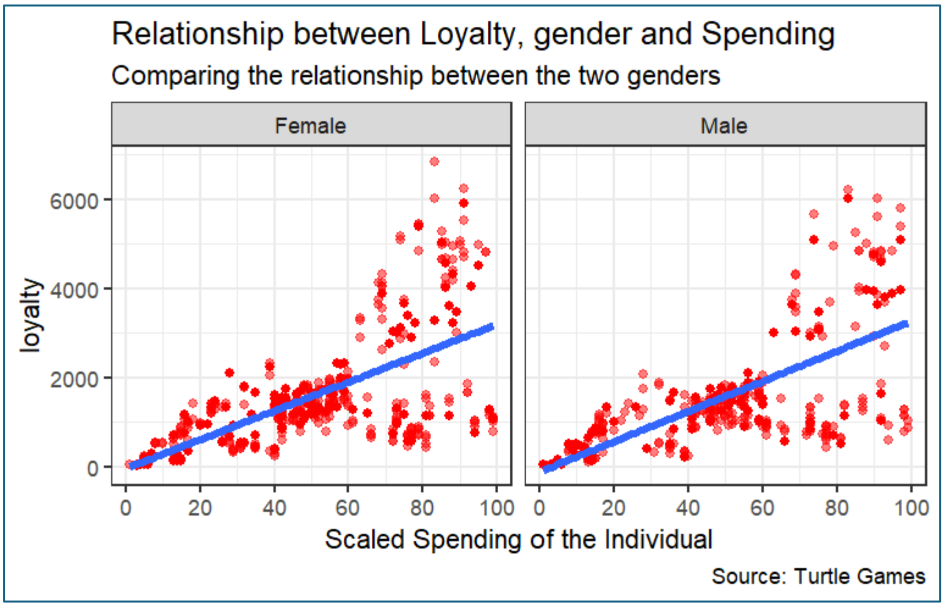
The 5 scatterplots show that salary has a positive impact on loyalty and most of the datapoints belong to graduate customers.

Figure 6.6: Scatterplots showing relationship between Loyalty Points, Spending and Education.



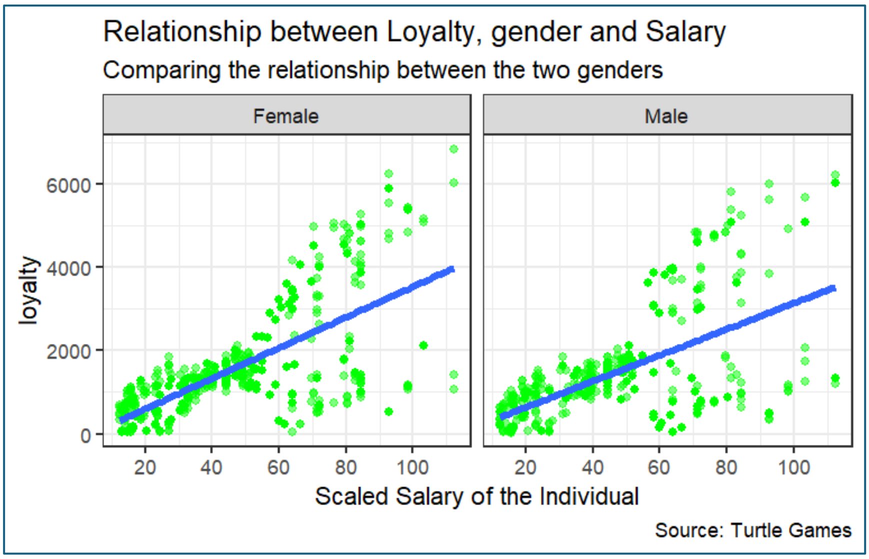
The 5 scatterplots show a similar trend that is seen in the previous visualisation. Spending also has a positive impact on loyalty and most of the datapoints belong to graduate customers.

Figure 6.7: Scatterplots showing relationship between Loyalty Points, Spending and Gender.



Females earn slightly more loyalty points than the male customers based on their spending score.

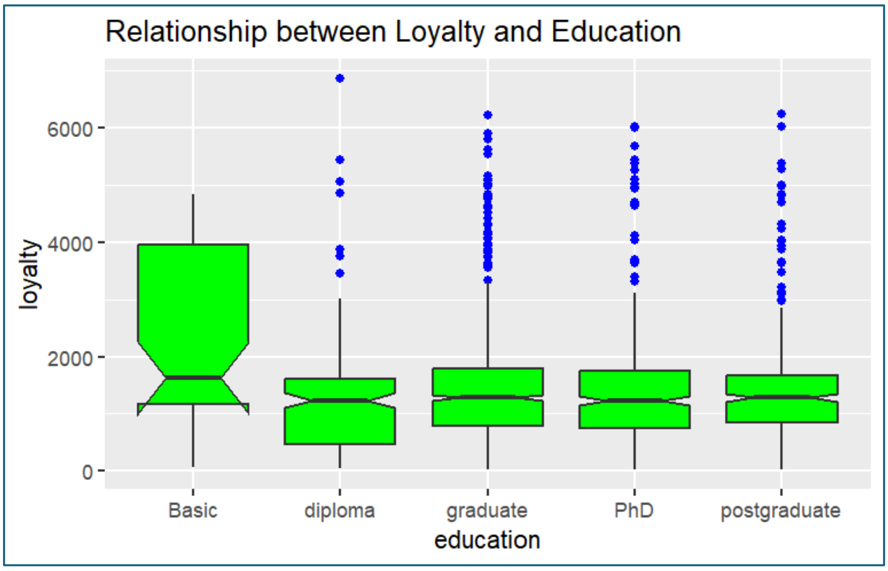
Figure 6.8: Scatterplot to indicate relation between Gender, Salary and Loyalty



The plot shows a positive relationship between salary and loyalty for both genders, meaning loyalty generally increases with higher salary. However, female customers exhibit a slightly steeper trend, suggesting that salary has a stronger influence on loyalty among women**.** Both groups show variability, but the pattern is more pronounced for females, indicating greater responsiveness to income in loyalty behaviour.

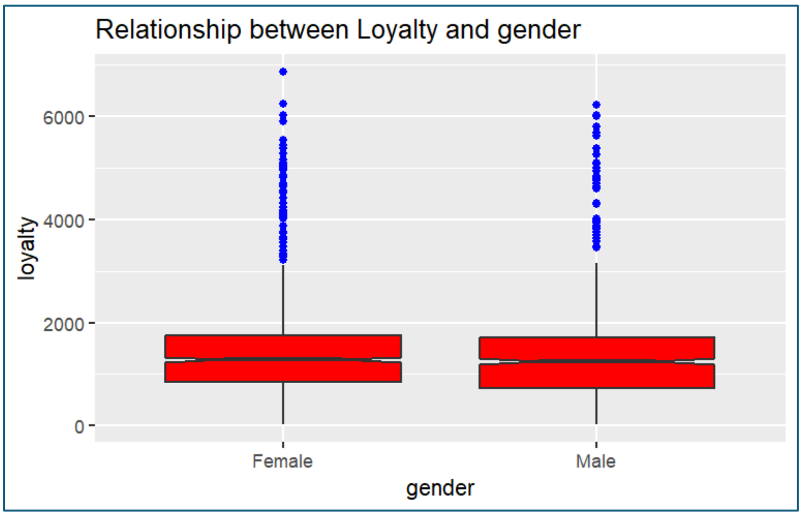
Appendix 7: Boxplots

Figure 7.1: Boxplot showing relationship between Loyalty and Education



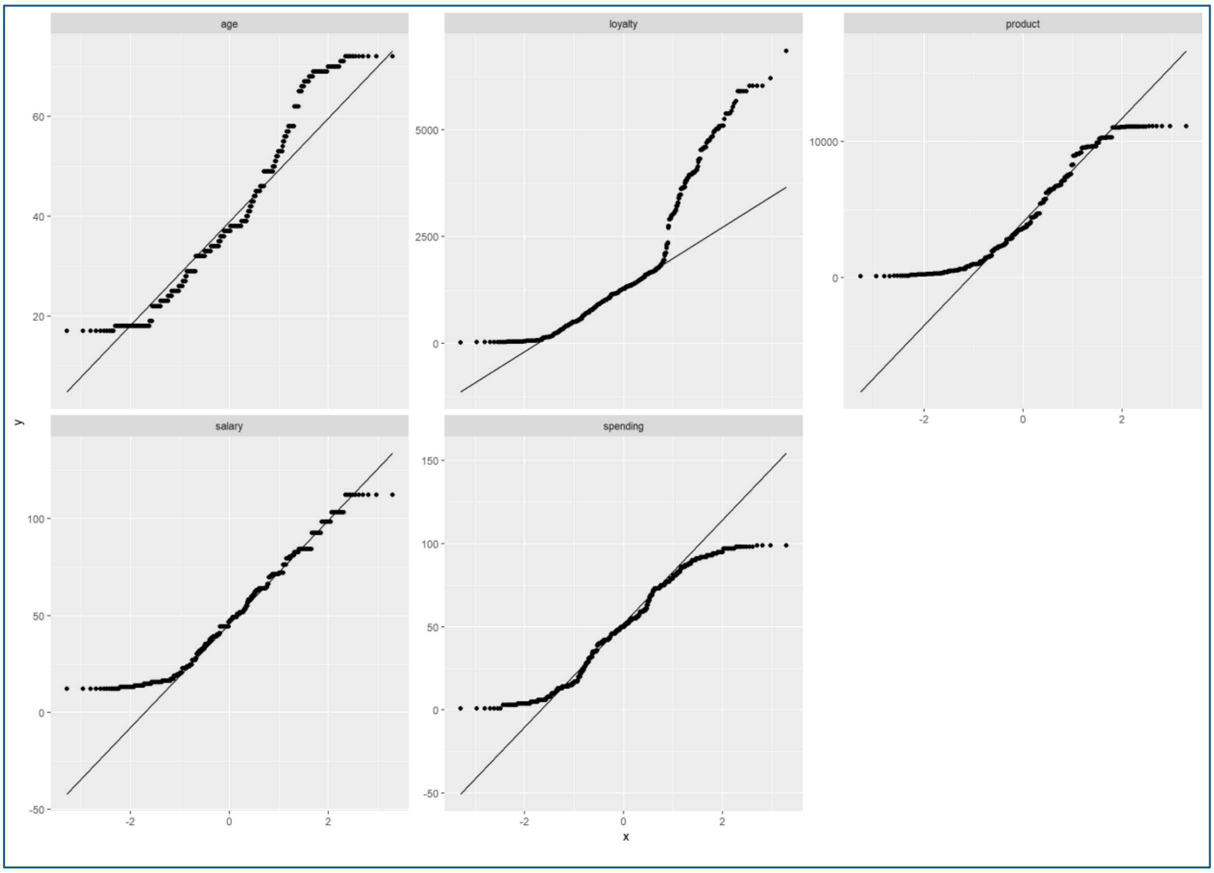
Customers with basic education show the highest average loyalty and the widest variability, suggesting they are generally more loyal but also diverse in behaviour. In contrast, customers with higher education levels (diploma, graduate, PhD, postgraduate) tend to have lower average loyalty, though some individuals in these groups are still highly loyal outliers. This indicates potential for targeted engagement among high-educated subgroups.

Figure 7.2: Boxplot showing relationship between Loyalty and Gender



The boxplot shows that loyalty scores are fairly similar between male and female customers, with nearly identical medians and interquartile ranges. Both groups have a comparable spread of data, though there are several high-loyalty outliers in each. Overall, gender does not appear to be a strong differentiator in loyalty levels based on this visual comparison

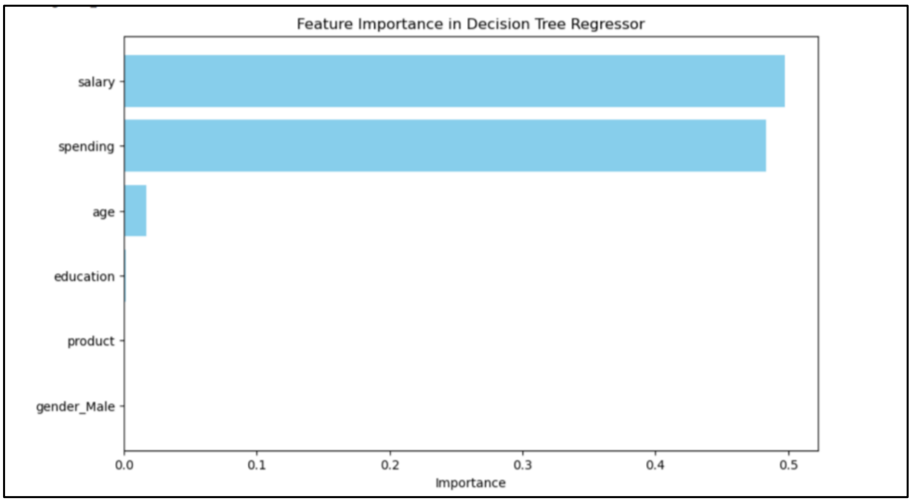
Appendix 8: QQ Plots showing Normality



The age and salary distributions are slightly right skewed, meaning a few higher values pull the average above the median. Both are close to normal in shape but still fail the Shapiro-Wilk test for normality. Spending is almost symmetric and has a flat distribution, with the mean and median around 50, but it's not perfectly normal either. Loyalty is heavily right skewed with many low values and some very high outliers, making it the least normally distributed variable. All four variables show some level of non-normality, especially loyalty, which may affect models that assume a normal distribution.

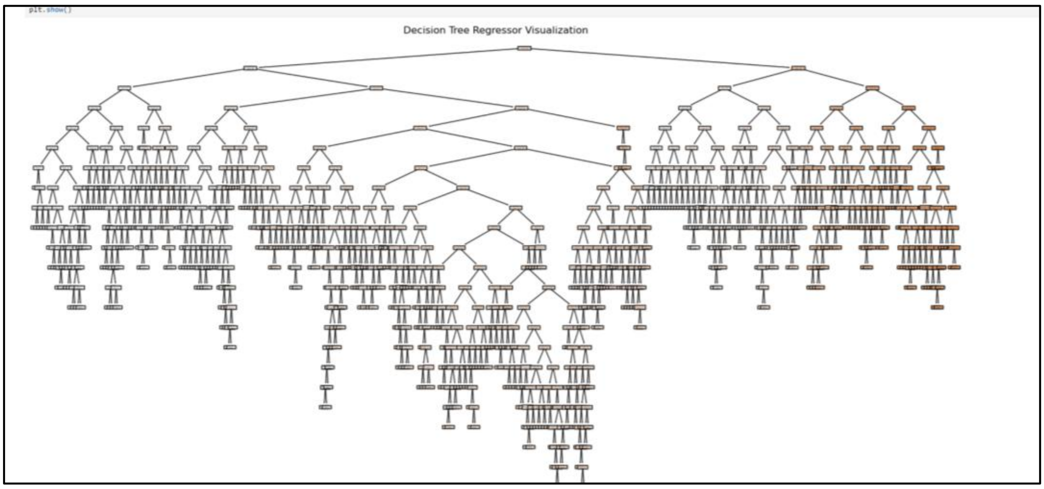
Appendix 9: Decision Tree Regressor

Figure 9.1: Feature Importance



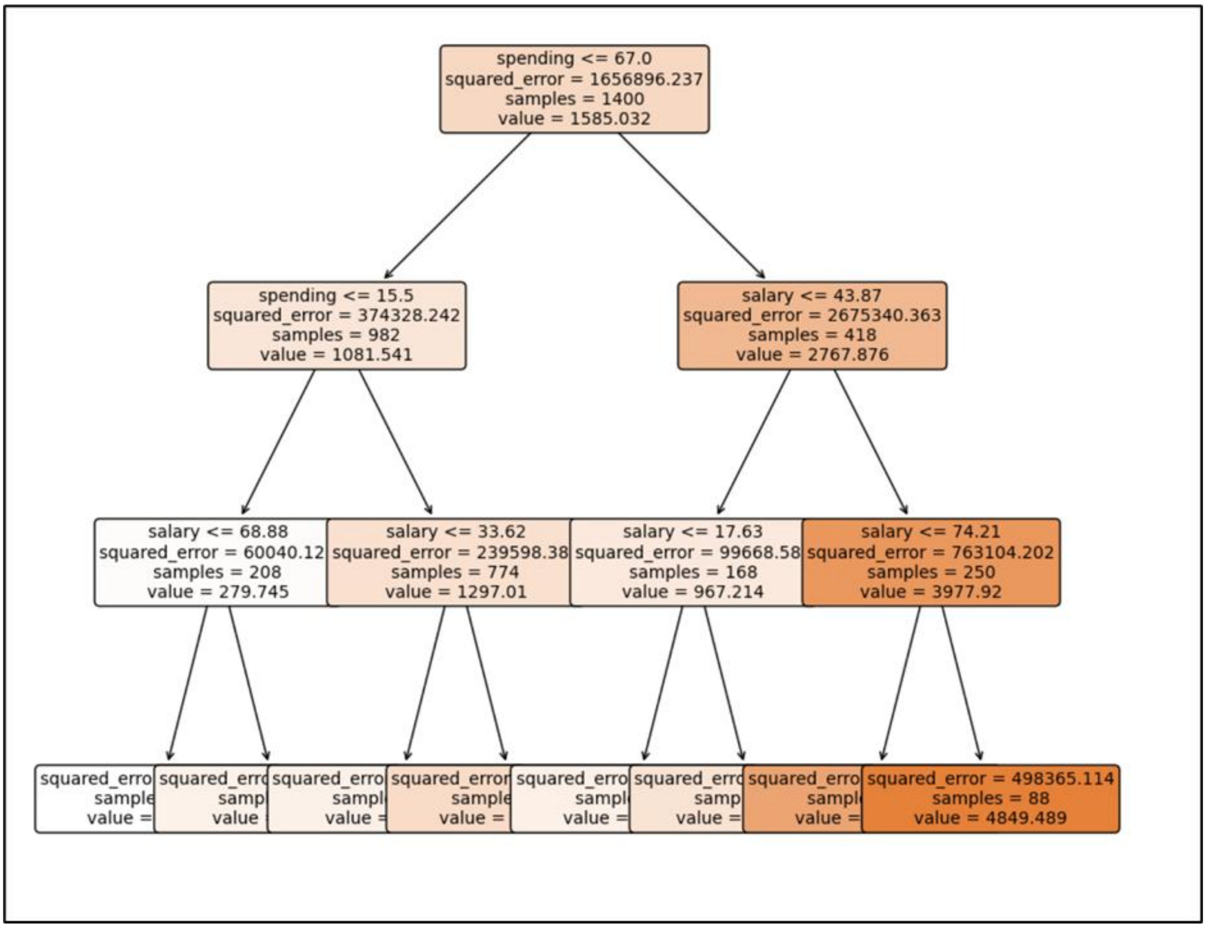
The visualisation shows that 'spending' and 'salary' have the most impact while 'age' has some impact on loyalty. Therefore, I will fit the decision tree regressor model on two different X data combinations and validate the accuracy of the model by checking the MAE, MSE RMSE before fitting the trained model on the test data.

Figure 9.2: The first Decision Tree Regressor Model.



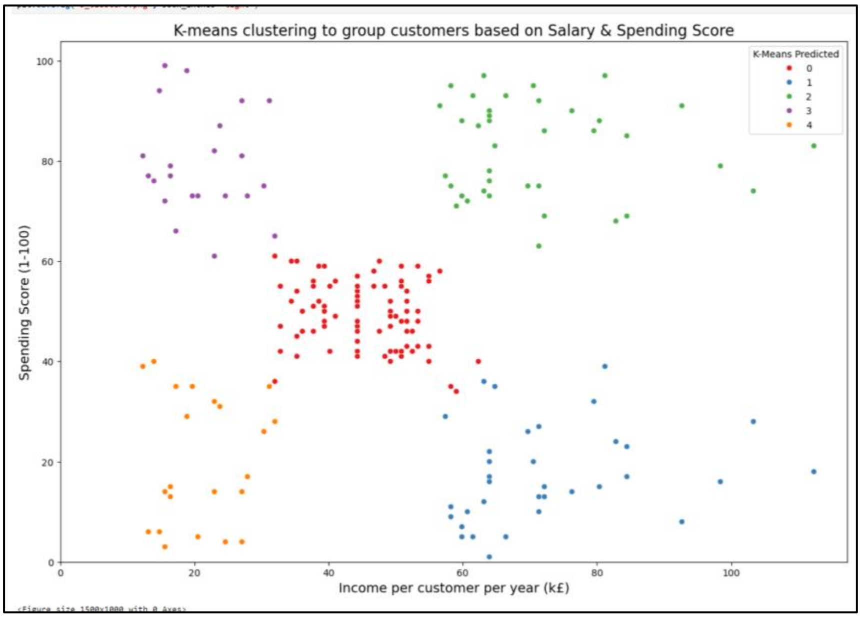
The model was too complicated to interpret. It is also prone to overfitting capturing noise instead of patterns.

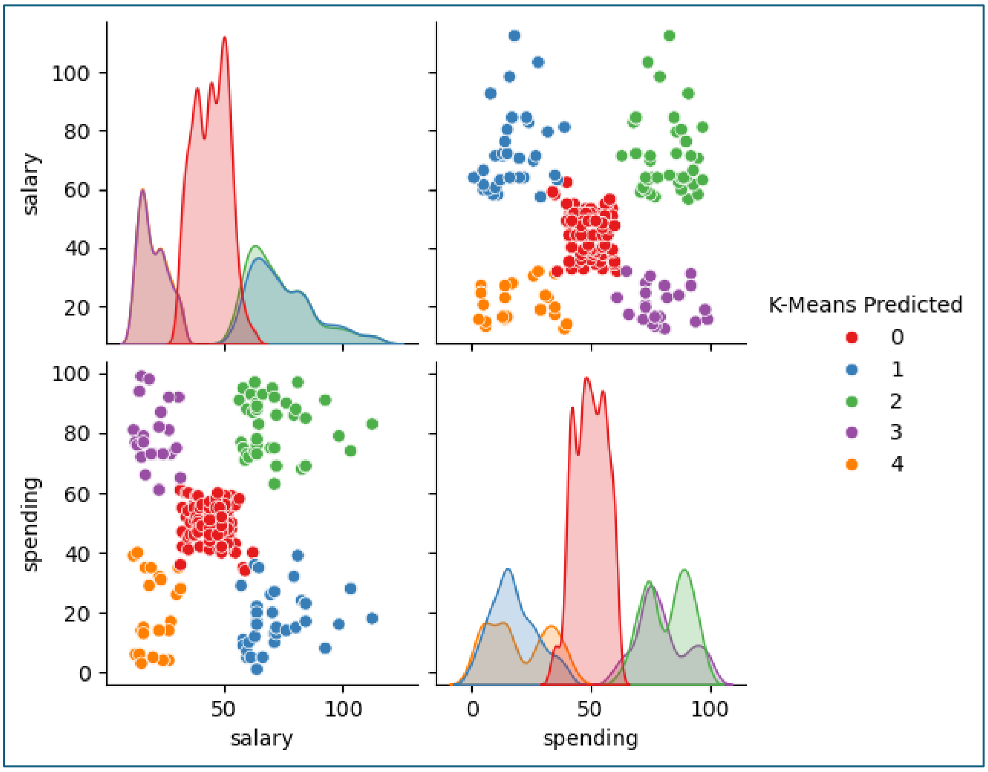
Figure 9.3: The second Decision Tree Regressor Model.



Pruning (limiting tree depth to 3 or 4) improved model simplicity but slightly worsened prediction error (MAE increased from 26 to 267) which means that the average predictions of the pruned model will be off by 267 units than the pre pruned model. There is no change in the R-Squared value (0.9961). Which means that it may not capture enough details or trends in the data. Despite this, the pruned tree was more interpretable, making it better for business use cases. The tree structure showed that **spending** was the most important variable (root node), followed by **salary**, and to a lesser extent, **age**. Splits such as spending <= 67 and salary <= 68.88 were key decision points.

Appendix 10: K-Means Clustering





The final model is the one with 5 clusters as the Elbow and Silhouette Methods graphs show 5 distinct hard clusters with no overlapping data points. The scatterplot and the pairplot confirm the above observation as well. For the purpose of customer segmentation, customers can be grouped in 5 clusters or categories based on their spending score and salary.

The Customer Segmentation – 5 Clusters:

- Cluster 0 (Red) belongs to customers with moderate income (30-60K£) and moderate spending score (30 - 60). They represent average shoppers who like to spend within their means.
- Cluster 1 (Blue) belongs customers with high income bracket (60-100K£) and low spending score (10-40). These customers should be engaged with offers.
- Cluster 2 (Green) belongs to high earners (60-100K£) with high spending score (Over 100). These customers should be offered premium offers.
- Cluster 3 (Purple) belongs low earners (20-40K£) with high spending score (60 -100). These are loyal customers who are perhaps keen to buy certain products.
- Cluster 4 (Yellow) belongs to customers with low income (20-40K£) and low spending score (Below 40). These customers are restricted by budget limitations.

Appendix 11: Customer Reviews Analysis

Figure 11.1: Word Cloud

Word Cloud showing the most frequently used words in Customers' Reviews Summary.

Figure 11.2: 15 Most Frequent Words occurring in Customers' Reviews.

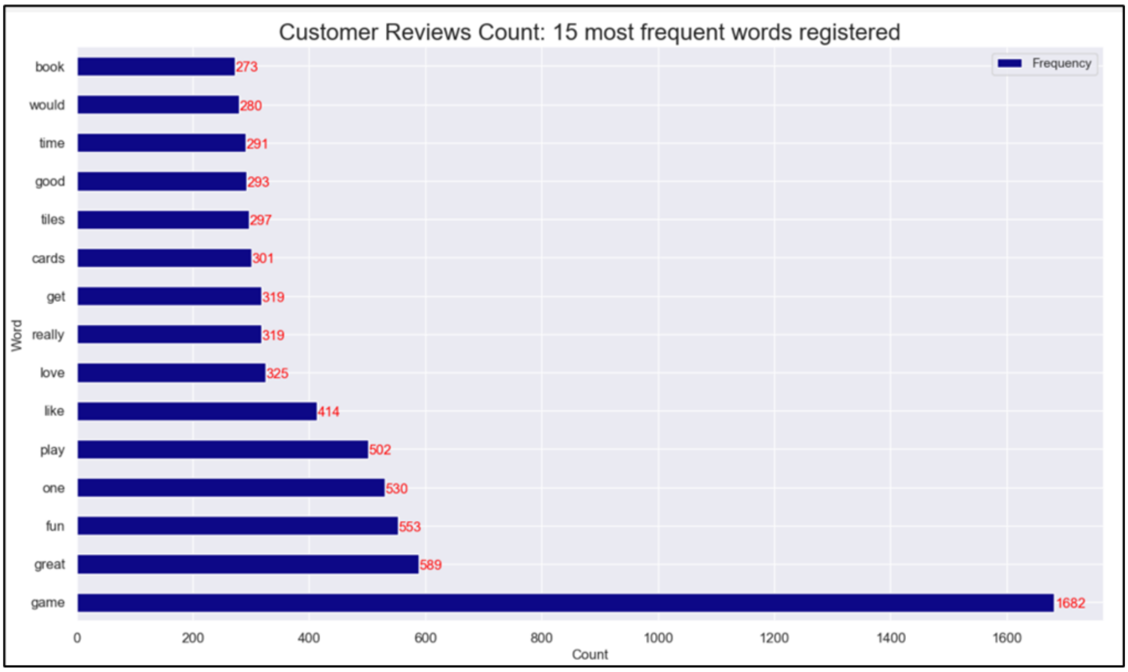
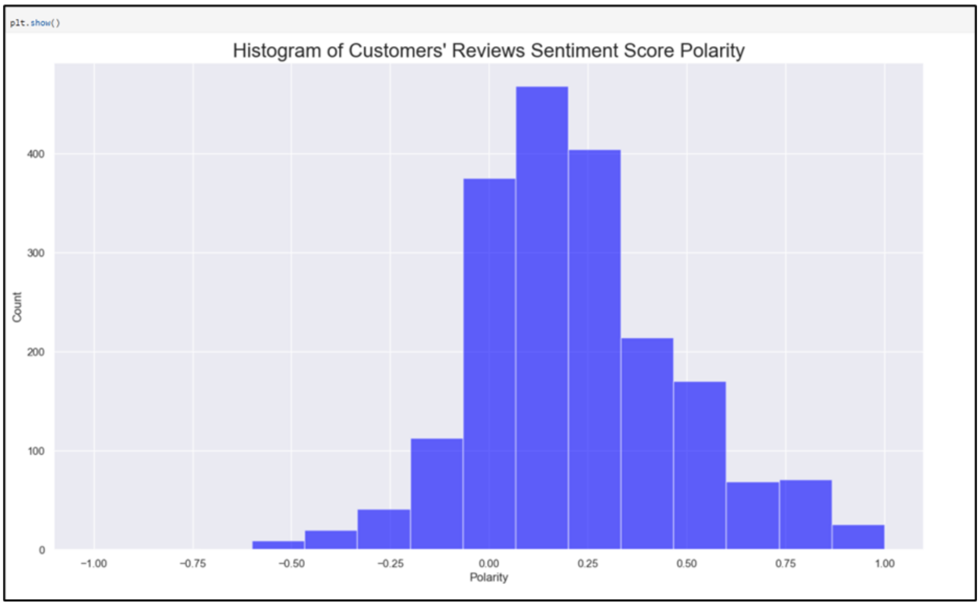
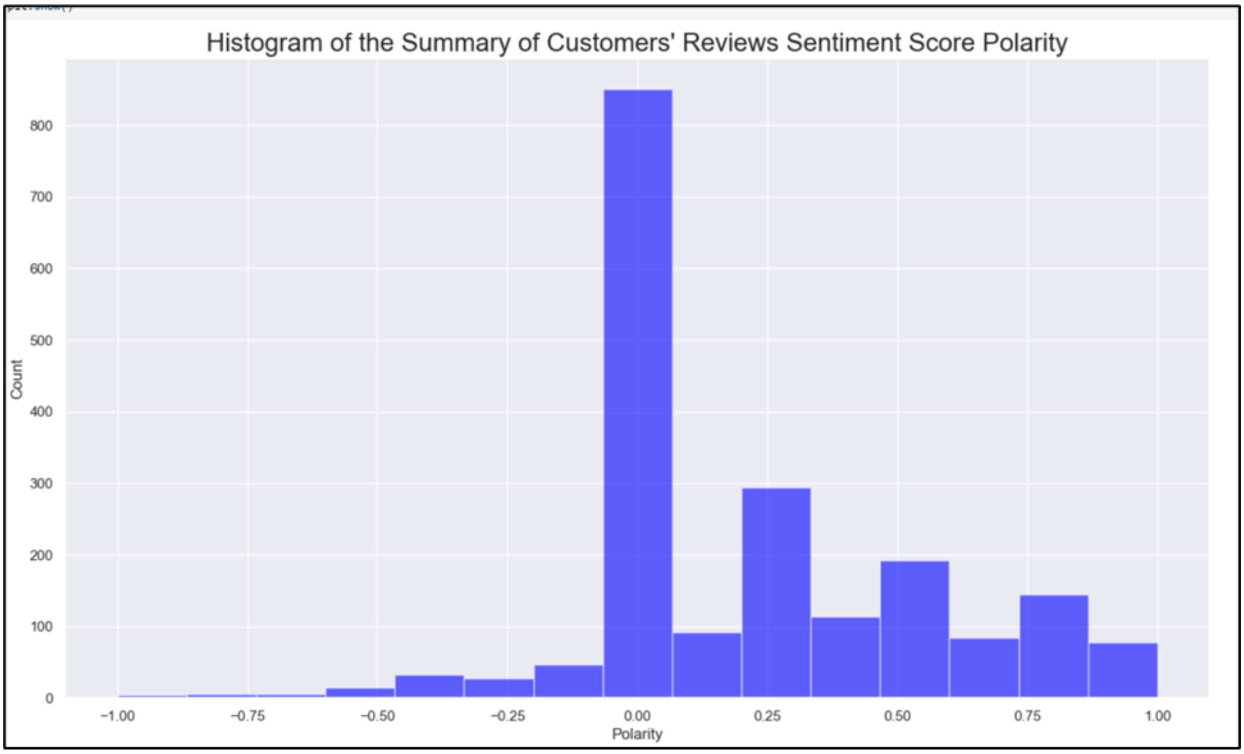


Figure 11.3: Histogram of Customer's Reviews Sentiment Score Polarity



This histogram shows that customers' reviews polarity score sits closest to neutral with a slightly stronger positive than the negative sentiment.

Figure 11.4: Histogram of Summary of Customers' Reviews Sentiment Score Polarity


Histogram of the Summary of Customers' Reviews Sentiment Score Polarity
Count
800
700
600
500
400
300
200
100
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Polarity

Appendix 12: Top 20 Positive and Negative Reviews and Summaries

Figure 12.1: Top 20 Positive Reviews

	review	review_polarity
7	came in perfect condition	1.000000
165	awesome book	1.000000
194	awesome gift	1.000000
492	excellent activity for teaching selfmanagement skills	1.000000
520	perfect just what i ordered	1.000000
587	wonderful product	1.000000
605	delightful product	1.000000
617	wonderful for my grandson to learn the resurrection story	1.000000
786	perfect	1.000000
928	awesome	1.000000
1127	awesome set	1.000000
1158	best set buy 2 if you have the means	1.000000
1167	awesome addition to my rpg gm system	1.000000
1290	its awesome	1.000000
1389	one of the best board games i played in along time	1.000000
1535	my daughter loves her stickers awesome seller thank you	1.000000
1593	this was perfect to go with the 7 bean bags i just wish they were not separate orders	1.000000
1697	awesome toy	1.000000
1702	it is the best thing to play with and also mind blowing in some ways	1.000000
1708	excellent toy to simulate thought	1.000000

The reviews are overwhelmingly positive customer feedback about various products — likely toys, books, or board games.

Figure 12.2: Top 20 Negative Reviews:

	review	review_polarity
208	booo unles you are patient know how to measure i didnt have the patience neither did my daughter boring unless you are a craft person which i am not	-1.000000
182	incomplete kit very disappointing	-0.780000
1786	im sorry i just find this product to be boring and to be frank juvenile	-0.583333
363	one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.550000
117	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000
227	this was a gift for my daughter i found it difficult to use	-0.500000
230	i found the directions difficult	-0.500000
290	instructions are complicated to follow	-0.500000
301	difficult	-0.500000
1511	expensive for what you get	-0.500000
174	i sent this product to my granddaughter the pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms the kit is useless since this was sent as a gift i do not have it to return very disappointed	-0.491667
346	my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	-0.446250
534	i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	-0.440741
306	very hard complicated to make these	-0.439583
423	kids i work with like this game	-0.400000
433	this game although it appears to be like uno and have an easier play method it was still too time consuming and wordy for my children with learning disabilities	-0.400000
493	my son loves playing this game it was recommended by a counselor at school that works with him	-0.400000
799	this game is a blast	-0.400000
802	i bought this for my son he loves this game	-0.400000
819	was a gift for my son he loves the game	-0.400000

References

Ankita. (2024, September 24). *K-Mean: Getting the Optimal Number of Clusters*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=The%20silhouette%20score%20is%20particularly%20helpful%20in%20determining, where%20data%20points%20are%20well-separated%20within%20their%20clusters.

Building a Decision Tree Regressor in Python: A Comprehensive Tutorial. (2023, 08 23). Retrieved from https://machinelearningtutorials.org/building-a-decision-tree-regressor-in-python-a-comprehensive-tutorial/

Detect and Remove Outliers using Python. (2024, August 30). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/

(Building a Decision Tree Regressor in Python: A Comprehensive Tutorial, 2023) (Detect and Remove Outliers using Python, 2024)